April 1962 • Vol. 16, No. 78

Mathematics of Computation

A journal devoted to advances in numerical analysis, the application of computational methods, mathematical tables, high-speed calculators and other aids to computation

NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
FOUNDED 1863

Formerly: Mathematical Tables and other Aids to Computation

Published Quarterly for the
National Academy of Sciences—National Research Council
By the
American Mathematical Society

Information to Subscribers

The journal is published quarterly in one volume per year with issues numbered serially since Volume I, Number 1. Starting with January, 1959 subscriptions are $8.00 per year, single copies $2.50. Other issues are available as follows:

Volume I (1943–1945), Nos. 10 and 12 *only* are available; $1.00 per issue.
Volume II (1946–1947), Nos. 13, 14, 17, 18, 19, and 20 *only* available; $1.00 per issue.
Volume III (1948–1949), Nos. 21–28 available. $4.00 per year (four issues), $1.25 per issue.
Volume IV–XII (1950 through 1958), all issues available; $5.00 per year, $1.50 per issue.

Microcard Edition

Volumes I–X (1943–1956), Nos. 1–56 are now available on Microcards and may be purchased from the Microcard Foundation, Box 2145, Madison 5, Wisconsin, at a cost of $20.00 for the complete set. Succeeding volumes are available on request.

Information to Contributors

All contributions intended for publication in *Mathematics of Computation* and all books for review should be addressed to H. Polachek, Technical Director, Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C. The author may suggest an appropriate editor for his paper. Manuscripts should be typewritten double-spaced in the format used by the journal. For journal abbreviations, see *Mathematical Reviews*, v. 21, Index, 1960. Authors should submit the original and one copy, and should retain one copy.

Subscriptions, address changes, business communications and payments should be sent to:

American Mathematical Society
190 Hope Street
Providence 6, Rhode Island

Published Quarterly for the
NATIONAL ACADEMY OF SCIENCES—NATIONAL RESEARCH COUNCIL
By the
AMERICAN MATHEMATICAL SOCIETY
Baltimore, Maryland and Providence, Rhode Island

del

sin,

sin,

red
are

.00

es),

.50

be
t a
t.

all
ied
The
be
via-
rig-

be

fe
fi
t
w
C

w
t
c
c
t

f
k
s
f
c
u

c
c
r

s
t

A Numerical Study of the Relative Class Numbers of Real Quadratic Integral Domains

By Harvey Cohn

1. Introduction. In a classic paper in 1856 Dirichlet gave some applications of a formula for the ratio of the class number of a quadratic integral domain in a real field to the class number of the whole integral domain (of all quadratic integers in that field), with the principal objective of showing that this ratio takes many values (such as 1) infinitely often for the real case, in support of a conjecture of Gauss.

The object of this paper is first of all to give Dirichlet's results briefly, together with some theorems and illustrations immediately deducible from them (in order to restrict the computation to cases in which the theory is of more help). We shall, of course, offer various tables of relative class numbers, such data being our main object. We emphasize quadratic integral domains of *prime power* conductor under the whole integral domain (of all quadratic integers of the field).

We ask, in particular, when the relative class number is divisible by 2 and 4, and find simple linear congruence conditions. When we ask which prime conductors have relative class numbers divisible by 3, we find such primes are essentially the splitting primes of certain cubic fields and therefore representable by quadratic forms, according to the classic work of Dedekind [3]. This is basically an application of class-field theory and perhaps the tables emerging would be of some experimental use. The classic background is amplified in [7], [5], and [2].

Here it might be appropriate to remark that the tables given below have a "natural" limit of diminishing returns owing to the fact that the relevant portions of classical algebraic number theory were developed long ago with relatively little data, and it would be desirable to see the theory profit from more data before great feats of computer endurance are attempted.

2. Notation and Terminology. We follow the convention that Latin letters generally denote rational integers and Greek letters denote algebraic integers. The following symbols and terms appear throughout the work:

m is a square-free integer > 1.

$R(m^{1/2})$ is the *field* generated by $m^{1/2}$.

d is the *discriminant* of the field generated by $m^{1/2}$; $d = m$ if $m \equiv 1 \pmod 4$, $d = 4m$ if $m \not\equiv 1 \pmod 4$.

c is the indicator of the type of field, $c = 2$ if $m \equiv 1 \pmod 4$, $c = 1$ if $m \not\equiv 1 \pmod 4$. Thus $d = 4m/c^2$.

$\mathfrak{O}$ is the set of all algebraic integers of $R(m^{1/2})$. It consists of $\omega = (x + ym^{1/2})/c$ for which x and y are rational integers subject only to the condition $x \equiv y \pmod c$.

Received August 18, 1961. This research was supported in part by a National Science Foundation grant. Computer service on GEORGE was donated by the Argonne National Laboratories of the U. S. Atomic Energy Commission through the cooperation of Dr. William F. Miller, Director of the Applied Mathematics Division.

$F(\omega)$ is the function defined by $y = F(\omega)$ in the above definition.

$\mathfrak{O}_f$ is an arbitrary integral domain (ring with unity) in $R(m^{1/2})$, given uniquely for any integer $f > 0$. It consists of the subset of algebraic integers ω in $\mathfrak{O}$ for which $f \mid F(\omega)$. Here f is called the *conductor*. It is the index of $\mathfrak{O}_f$ in $\mathfrak{O}$, and $\mathfrak{O}_1 = \mathfrak{O}$.

f^2d is the *ring-discriminant* of $\mathfrak{O}_f$. Its purpose is that any given $D(>1)$ which is $\equiv 0$ or $1 \pmod 4$ can be written uniquely as f^2d for some f and d. Thus f^2d completely determines $R(m^{1/2})$ and $\mathfrak{O}_f$ in $R(m^{1/2})$.

$h(f^2d)$ is the class number (of ideals prime to f) in $\mathfrak{O}_f$.

$H(f)$ is the relative class number $= h(f^2d)/h(d)$, (used when the value of d, or the field, is understood in context).

ϵ is the fundamental unit, written $\epsilon = (a + bm^{1/2})/c$. Here $a > 0$, $b > 0$ and $a \equiv b \pmod c$.

e is the norm of the fundamental unit, actually ± 1, $N(\epsilon) = e = (a^2 - mb^2)/c^2$.

$\psi(f)$ is the value of $f\Pi(1 - (d/q)/q)$ extended over primes q which divide f. Here (d/q) is the Kronecker residue symbol, (thus $(d/2) = (2/d)$).

$\phi(f)$ is the minimum exponent t (>1) for which $\epsilon^t \varepsilon \mathfrak{O}_f$ or for which $f \mid F(\epsilon^t)$. It can be shown directly that $\phi(f) \mid \psi(f)$. By classical methods of primitive root theory, if $f \mid F(\epsilon^u)$, then $\phi(f) \mid u$.

Other symbols appear only locally and can best be defined as they arise.

3. Dirichlet's Theorems. The starting point is the following theorem, in principle due to Gauss: *For a given field* $R(m^{1/2})$, (*with* $m > 0$),

$$H(f) = \psi(f)/\phi(f). \tag{3.1}$$

If $m < 0$, the formula is modified so that, for instance, with $f > 1$, $\phi(f)$ is replaced by half the number of units in $\mathfrak{O}$. (We do not need the modified formula for the machine part of the calculation, but for supporting computations in Section 7).

Now $\psi(f)$ is fairly easy to find, but the calculation of $\phi(f)$ is the part requiring the electronic computer. Dirichlet [4] showed, however, that if $f = p_1^{F_1} \cdots p_s^{F_s}$, where the primes p_i come from a given finite set, then the values of $H(f)$ also come from a finite set as the exponents F_i vary; in fact $H(f) = H_0$, a constant if each F_i is sufficiently large. An examination of Dirichlet's method leads to the rule that if p_i is odd and f_0 is such that $H(f_0) = H(f_0p_i)$ (while if one $p_i = 2$, f_0 satisfies $H(f_0) = H(4f_0)$), then $H(f) = H(f_0)$ if $f_0 \mid f$, (recalling the prime divisors of f are to be limited to the p_i).

From general principles it also follows that if $f \mid g$, then $H(f) \mid H(g)$.

The main step in understanding these results is to consider any f which contains all the odd primes p_i (and possibly 2^2) as divisors. Then $f \mid F(\epsilon^{\phi(f)})$, i.e., $\epsilon^{\phi(f)} = (x_f + y_fm^{1/2})/c$, where $f \mid y_f$. But, let f^* be the factor of y_f consisting of powers of the p_i. (Thus $f \mid f^*$ while $(y_f/f^*, f) = 1$.) Then for p_i odd, $F(\epsilon^{\phi(f)p_i}) = p_if^*g$, where $(g, f) = 1$, as we prove by using the binomial theorem, (in a manner reminiscent of the proof that a primitive root modulo p^2 is a primitive root modulo p^n, $n > 2$). For p_i even, special attention must be given the denominator $c = 2$, but this can be left to the reader, as well as the completion of the proof of the above results by induction.

If we restrict f to powers of a prime p then we find $H(p^{n+1}) = H(p^n)$ or $pH(p^n)$ $(n \geqq 1)$, but eventually $H(p^{n+1}) = H(p^n)$ then $H(p^m) = H(p^n)$ for all $m \geqq n$ when p is odd, while for $p = 2$, $H(2^{n+1}) = H(2^n)$ or $2H(2^n)$, $(n \geqq 1)$, but eventually $H(2^{n+2}) = H(2^n)$ where $H(2^m) = H(2^n)$ for all $n \geqq m$.

4. Simple Cases. We first consider those q which divide 6 mb. In these cases the values of $H(q^t)$ are easily seen by elementary hand calculations, and we often omit these from the tables to make room for more interesting values.

$$\mathbf{f = 2^F}$$

$$\text{Define } M(a, b, f) = \begin{cases} 1 & \text{if} \quad f = 1, \\ \min\,(2^B, f) & \text{if} \quad 2^A = 1, \qquad f \geqq 2, \\ \min\,(2^A, f/2) & \text{if} \quad 2^A > 1, \qquad f \geqq 2, \end{cases}$$

where $2^A \parallel a$, i.e., $2^A \mid a$ but $2^{A+1} \nmid a$, and likewise $2^B \parallel b$. Then if $d \equiv 0 \pmod 4$,

$$H(f) = M(a, b, f), \tag{4.1}$$

while if $d \equiv 1 \pmod 4$ and $2 \mid ab$,

$$H(f) = [2 + (d/2)]M(a/2, b/2, f/2), \tag{4.2}$$

and if $d \equiv 1 \pmod 4$ and $2 \nmid ab$ (whence $d \equiv 5 \pmod 8$)),

$$H(f) = M([a^2 - 3e]/2, [a^2 - e]/2, f/2). \tag{4.3}$$

(Note that $([a + bm^{1/2}]/2)^3 = a[a^2 - 3e]/2 + b[a^2 - e]m^{1/2}/2$).

$$\mathbf{f = 3^F}$$

Let $3^B \parallel b$, $3^A \parallel a$. If $3 \mid m$, let $3^G \parallel 3a^2 + mb^2$, then

$$H(f) = \begin{cases} \tfrac{1}{3} \min\,(f, 3^G) & \text{if} \quad 3^B = 1 \\ \min\,(f, 3^B) & \text{if} \quad 3^B > 1. \end{cases} \tag{4.4}$$

If, however, $3 \nmid m$, let $3^T \parallel a^2 + b^2 m$, then

$$H(f) = \begin{cases} \tfrac{1}{3} \min\,(f, 3^T) & \text{if} \quad 3 \nmid ab, \\ \tfrac{1}{2}[1 - (d/3)/3] \min\,(f, 3^A) & \text{if} \quad 3 \mid a, \\ [1 - (d/3)/3] \min\,(f, 3^B) & \text{if} \quad 3 \mid b. \end{cases} \tag{4.5}$$

(Note that $f = 3^F$ is "special" because of consideration of 3^G. Compare $f = q^F$ below).

$$\mathbf{f = q^F}$$

Here let q be a prime $\neq 2, 3$ for which $q \mid mb$, and let $q^B \parallel b$. Then

$$H(f) = \min\,(f, q^B). \tag{4.6}$$

Thus in many cases where $q \mid m$ and $q \nmid 6b$, then $H(q^n) = 1$ for all n, giving the easiest illustration of Dirichlet's original objective; e.g., for $m = d = 5$, $H(5^n) = 1$ for all $n > 0$.

5. The Program. The basic sub-routine considers the input

$$m, a, b, f \tag{5.1}$$

from which $\phi(f)$, $\psi(f)$, and $H(f)$ are calculated. The machine forms by induction $\epsilon^t = [a(t) + b(t)m^{1/2}]/c$ stored as $a(t)$, $b(t)$ calculated modulo f^2. Then letting $t = 1, 2$, the machine records the earliest $t[= \phi(f)]$ for which $b(t) \equiv 0 \pmod f$. The machine next calculates $\psi(f)$ by examining the prime factors q of f sequentially. The machine finds (d/q) for $q \nmid 2d$ by actually testing the solvability in x of $x^2 \equiv d \pmod q$, while for $q \mid 2d$, (d/q) is determined directly from the rules. Finally, $H(f) = \psi(f)/\phi(f)$. The output for each input consists of

$$f, H(f), \psi(f), b(\phi(f))/f \pmod f. \tag{5.2}$$

The last value is desired for purposes of testing $F(\epsilon^{\phi(f)})$. For example, if

$$(b(\phi(f))/f, f) = 1,$$

then f is a suitable f_0 for Section 3.

The basic sub-routine was used in several ways.

In one run the basic sub-routine was set up to increment f by 1 automatically over a range $f_1 \leqq f \leqq f_2$ where f_1 and f_2 are given in addition to the initial data. For $m = 5$ the problem was run up to $f = 4400$ and for $m = 2$ and 3, it was run up to $f = 1000$.

In another variation, the values of f were incremented as before but were restricted to *primes* in the preassigned range. (We always use the letter p to denote a prime.) These main runs were made for $f = p$ an odd prime up to 997 for 38 values of m, namely

$$\text{Series A: } 2 \leqq \text{square free } m \not\equiv 1 \pmod 4 \leqq 42 \tag{5.2}$$

$$\text{Series B: } 5 \leqq \text{square free } m \equiv 1 \pmod 4 \leqq 97. \tag{5.3}$$

The problem was programmed for the GEORGE computer with only approximately 500 words of a 4096-word high-speed memory involved. The machine is internally binary with 40-bit word length and approximate speed of 50,000 two-address operations per second.

In all the runs, the output consisted of the input data (5.1) (as a heading) followed by the output data (5.2) listed "on-line" (parallel) with the computation. The input and output were in decimal (internally converted) and on paper tape originally (but the output was later transformed to magnetic tape just to speed up the printing process from flexowriter to line printer). The actual input and output times were negligible.

The running time for each case was about $f/50$ seconds. The calculations were run between December 1960 and May 1961.

6. Use of Some Cyclic Groups. Let m be given and let $p \nmid 2m$ be an arbitrary given prime. Define a group in which the elements $\mathfrak{a}_t$ are the following sets:

$$\mathfrak{a}_t = \{x + ym^{1/2}\}, \quad \text{where} \quad x \equiv ty \quad \text{and} \quad N(x + ym^{1/2}) \not\equiv 0 \pmod p, \tag{6.1}$$

and

(6.2) $\mathfrak{a}_\infty = \{x\}$, where $x \neq 0$.

The group operation is multiplication (mod p), easily shown to be independent of the representative. When $(m/p) = -1$, there are $p + 1$ of these elements, while when $(m/p) = +1$ there are $p - 1$ of these elements (by excluding two values of t for which $t^2 \equiv m \pmod{p}$). In general, we have a group $\mathfrak{A}_p$ with $p - (m/p) = \psi(p)$ elements, and with $\mathfrak{a}_\infty$ as the unit element.

We see that the group $\mathfrak{A}_p$ is cyclic. This is true where $(m/p) = -1$ since the group is a sub-group of the cyclic group of reduced residues of algebraic integers modulo p, (now an ideal prime). When $(m/p) = -1$ we rewrite $\mathfrak{a}_t = \mathfrak{a}[u]$ where

$$(6.3) \qquad \mathfrak{a}[u] = \{x(r(1 + u) + m^{1/2}(1 - u))\}.$$

Here r satisfies $r^2 \equiv m \pmod{p}$ and t and u are related by $t \equiv r(1 + u)/(1 - u) \pmod{p}$. We can verify $\mathfrak{a}[u]\mathfrak{a}[v] = \mathfrak{a}[uv]$, hence when $(m/p) = 1$, $\mathfrak{A}_p$ is isomorphic to the multiplicative (cyclic) residue group of rational integers modulo p.

The important result for us is the following: if $p \nmid 2m$ and if r is a given integer dividing $p - (m/p)$ a necessary and sufficient condition that $r \mid H(p)$ is that $c\epsilon$ belong to an $\mathfrak{a}_t$ which is an r-th power in $\mathfrak{A}_p$. This result follows from the cyclic structure of $\mathfrak{A}_p$ once we note that $(c\epsilon)^{\phi(p)} \equiv z \pmod{p}$ for z an integer, hence $(c\epsilon)^{\phi(p)}$ belongs to $\mathfrak{a}_\infty$ the unit element, while $\psi(p)$ is the order of the group.

For illustration, we start with $r = 2$, and take $p \nmid 2mb$. Set

$$a + bm^{1/2} = (x + ym^{1/2})k, \quad \text{or,}$$

$$(6.4) \qquad \begin{cases} a \equiv k(x^2 + y^2m) \\ b \equiv 2kxy. \end{cases}$$

This system is solvable, for $k \neq 0$, if and only if the equation

$$(6.5) \qquad bx^2 - 2axy + bmy^2 \equiv 0 \bmod p$$

is solvable, with $(x, y) \neq (0, 0)$. The discriminant is $4c^2e$. Hence if $N(\epsilon) = e = -1$, then $2 \mid H(p)$, for $p \nmid 2mb$.

Thus for some cases, e.g., where $N(\epsilon) = +1$, the only possible f for which $H(f) = 1$ must come from primes in the special cases in Section 4 above. (We recall that if $f \mid g$, then $H(f) \mid H(g)$). Thus for $m = 3$, the only f for which $H(f) = 1$ are now seen to be $f = 3^t$ and $f = 2 \cdot 3^t$.

We next consider the sub-group of $\mathfrak{A}_p$, called $\mathfrak{B}_p$, all of whose elements have norms which are quadratic residues of p. Thus $\mathfrak{a}_\infty$ is necessarily in $\mathfrak{B}_p$, while $\mathfrak{a}_t$ is in $\mathfrak{B}_p$ if and only if $([t^2 - m]/p) = +1$. It is easily seen that the norms of representatives in $\mathfrak{a}_t$ are not all residues, by results on successions of residues and non-residues. Thus $\mathfrak{B}_p$ has only order $(p - (m/p))/2$, since it must then be of index 2. Now if we normalize the representative of $\mathfrak{a}_t$ in (6.1) belonging to $\mathfrak{B}_p$ to be plus or minus an element of norm 1, we can say that if $e = 1$, then ϵ represents a perfect square in $\mathfrak{B}_p$ if and only if for some integers x and y

$$(6.6) \qquad \pm c^2\epsilon \equiv (x + ym^{1/2})^2 \pmod{p}.$$

But the condition for a perfect square in $\mathfrak{B}_p$ is precisely the condition that $\pm\epsilon$ represents a perfect fourth power in $\mathfrak{A}_p$, or $4 \mid H(p)$. Expanding (6.6), we discover we must be able to solve simultaneously

$$(6.7)\qquad \begin{cases} \pm ca \equiv x^2 + my^2 \\ \pm cb \equiv 2xy \end{cases} \pmod{p}.$$

An elementary calculation reveals this system is solvable if and only if (with signs $s_1, s_2 = \pm 1$),

$$(6.8)\qquad \begin{cases} x^2 + my^2 \equiv s_1 ca \\ x^2 - my^2 \equiv s_2 \end{cases} \pmod{p}.$$

For this it is necessary and sufficient that $2c(s_1 a + s_2)$ and $2mc(s_1 a - s_2)$ be perfect squares modulo p. With some manipulation, we find, *if $N(\epsilon) = e = +1$ and $p \nmid 2mb$, then a necessary and sufficient condition that $4 \mid H(p)$ is that*

$$(6.9)\qquad (-1/p) = (m/p) = ([2a/c - 2]/p).$$

We can often simplify the result (6.9) to take the form

$$(6.10)\qquad (-S/p) = (Q/p) = (R/p),$$

for smaller values of Q and R shown in the columns 10 and 11 of Table I with $S = 1$, except for $m = 15$ and 35, where $S = 2$. When $e = -1$, there are still many occurrences of $H(p) = 4$ (the smallest such value is listed in column 11).

7. Divisibility by 3. A more interesting case is $r = 3$. This can occur (for $p \nmid 6mb$) only when $3 \mid \psi(p)$ or $(-3m/p) = 1$. We ask, when can we solve $c\epsilon \equiv k(x + ym^{1/2})^3 \pmod{p}$ or

$$(7.1)\qquad \begin{cases} a \equiv k[x^3 + 3xy^2m] \\ b \equiv k[3x^2y + y^3m] \end{cases} \pmod{p},$$

for $xy \not\equiv 0$? Eliminating k, we see this leads to the solvability of $\lambda(x/y) \equiv 0 \bmod p$ where λ is a polynomial defining a root of a cubic field,

$$(7.2)\qquad \lambda(\xi) = b\xi^3 - 3a\xi^2 + 3b\xi m - am = 0.$$

Hence $3 \mid H(p)$ (for $p \nmid 6m$) if and only if p is a splitting prime for the field $R(\xi)$. In fact, *p must split into three distinct prime ideals* since $(-3m/p) = 1$, and the discriminant D_3 of the cubic can be shown to differ from $-3m$ by a rational square. The reader is referred to Hasse's work [6] for details on the method.

Finding the field discriminant of $R(\xi)$ is rather lengthy but since the methods are so well-known we can merely outline the steps. The module $[1, b\xi, am/\xi]$ consists only of integers of $R(\xi)$ and its discriminant is $-108mc^4$ by a direct calculation. Since only perfect squares could be superfluous factors of the discriminant, we need examine the basis elements to see if $r + sb\xi + tam/\xi$ can be divisible by 2 (or 3) without r, s, and t being simultaneously divisible by 2 (or 3). We find the *only* possibilities are the following cases which we leave for the reader to verify:

Case i. $3 \mid m$ and $3 \mid b$; then $3 \mid b\xi$ and $3 \mid (am/\xi)$
Case ii. $3 \nmid m$ and $9 \mid a$ (or b); then $3 \mid b\xi$ (or $3 \mid (am/\xi)$)
Case iii. $3 \nmid mab$ and $am \equiv \pm b \pmod{9}$; then $3 \mid (b\xi + e_1e_2am/\xi - e_2)$

where $e_1 = \pm 1 \equiv am$, $e_2 = \pm 1 \equiv b \pmod{3}$

Case iv. $c = 2$; then $2 \mid b\xi$, $4 \mid (b\xi + am/\xi)$.

These calculations were made partly on the basis of possible ideal factorizations of (2) and (3) and partly as a direct consequence of the following equation for $\mu = (b\xi + am/\xi)e$:

$$\begin{aligned}\mu^3 - 3b(1 - m)\mu^2 &+ 3(b^2 - a^2)(1 - m)\mu \\ &+ [a^3(6m + 2) + a^2b(m^2 - 12m + 3) \\ &\qquad - ab^2(2m + 6m^2) + b^3(9m^2 - 1)] = 0.\end{aligned} \tag{7.3}$$

The occurrences of cases (i–iii) are noted in column 7 of Table I.

We finally obtain

$$d_3f_3^2 = D_3 = -108m/s^2c^2, \tag{7.4}$$

where

$$\begin{cases} s = 9 & \text{if } 3 \mid m, 3 \mid b, \\ s = 3 & \text{if } 3 \nmid m, 9 \mid ab, \quad \text{or if } 3 \nmid mab, am \equiv \pm b \pmod 9, \\ s = 1 & \text{otherwise.} \end{cases} \tag{7.5}$$

We then consider the set of $h(d_3f_3^2)$ primitive reduced quadratic forms of discriminant D_3. Those which are perfect cubes under composition represent precisely all primes $p(\nmid 6m)$ for which $3 \mid H(p)$.

A supporting computation was made by Mr. Roy Lippmann on an IBM 650 to calculate all primitive reduced forms from D_3. The square-free kernel m_3 is shown in Table I, together with $h(D_3)$ and the conductor f_3. The $h(D_3)$ primitive forms (A, B, C) which are cubes under composition were most easily identified by finding some "convenient" small prime $(p \nmid 6m)$ represented by the form and checking $H(p)$, (see [1]). The coefficients A and B of forms and representative primes p and $H(p)$ are listed in Table III.

Now in every case, it so happens that $3 \parallel h(d_3f_3^2)$, hence there are $h(d_3f_3^2)/3$ forms which are perfect cubes. Also, the ambiguous forms are always perfect cubes, but in general they are not the complete set. The non-ambiguous forms, naturally, are written two at a time by means of $\pm B$.

8. Irregular Primes. We finally note that there are many odd primes p, for which, for some fixed $i > 0$,

$$H(p^n) = H(p) \min (p^{n-1}, p^i). \tag{8.1}$$

We call these primes irregular and we call i the *index of irregularity*. When $p \nmid 6mb$ such cases are explained by some combinational curiosities much less transparent than those occurring in Section 3. They are listed because the occurrence of prime divisors of f in the relative class number is of some theoretical value.

These values were found by scanning the outputs (5.2) as f ran over the odd primes p for cases where $b \equiv 0 \pmod p$. The 53 individual cases which emerged were tested by rerunning these cases, using $f = p^2$. The values of $b/f \not\equiv 0 \pmod f$ indicated primes of index 1, while those where $b/f \equiv 0$ while $b/fp \not\equiv 0 \pmod f$ indicated primes p of index 2. No odd primes of higher index emerged from the experiment.

9. Summary of Calculations. The problem ran some 40 hours and generated some 300 pages of tables—obviously too much to reproduce! We therefore attempt a qualitative résume.

From the output, we would readily believe that when $e = -1$ there are infinitely many odd primes for which $H(p) = 1$, while when $e = 1$ there are infinely many primes for which $H(p) = 2$. Indeed, even in the case $e = 1$, we know (from Section 4) that if $p \mid m$ and $p \nmid 6b$ then $H(p) = 1$. In either case, except for scattered irregular primes in Table IV, $H(p) = H(p^n)$.

A frequency count is surprising in its uniformity. When $e = -1$, we examine the 167 odd primes < 1000 and find $H(p) = 1$ in 39–43 per cent of these primes as m varies, while when $e = +1$ the corresponding case $H(p) = 2$ occurs for 56–63 per cent of these primes as m varies. If we define $P(m, n; x)$ as the proportion of primes $\leqq x$ for which $H(p) = n$ (in reference to $R(m^{1/2})$) we find a reasonably steady value for $P(5, 1; x)$. For instance, $P(5, 1; 500) = 42$ per cent, $P(5, 1; 1000) = 41$ per cent, $P(5, 1; 2000) = 39$ per cent, $P(5, 1; 4000) = 37$ per cent.

Continuing with $m = 5$, $H(p)$ (as far as we might imagine) "should" take all prime values but it seems to take large values "rather slowly." The earliest p for some larger primes are $H(911) = 13$, $H(1087) = 17$, $H(3079) = 19$, $H(1103) = 23$. For $p < 4400$, $H(p)$ takes no larger prime! Thus an "asymptotic" study of the values of $H(p)$ can be expected to be astronomical in size (perhaps larger than for studies of classical prime number distributions).

Table II is given to point out some relative class numbers which are small prime powers; $H(p) = 3$ is in Table III; and $H(p) = 2$ or 4 comes from columns 10 and 11 of Table I. Despite the uniformity of the earlier frequency count, some values of m seem to be more "amenable" to given values of $H(p)$ than others. This seeming paradox might again be a manifestation of the fact that "$p < 1000$" is a miniscule range of values!

As far as *congruence* properties of $H(p)$ are concerned, Sections 6 and 7 provide us with much more guidance. For example, by the uniform density of primes in linear congruence classes for a fixed modulus, when $e = -1$, $H(p) \equiv 0 \pmod 4$ only one-third as often as $H(p) \not\equiv 0 \pmod 4$.

In a similar manner, using known results on the distribution of primes represented by quadratic forms [8], we can see that if k_3 of the $h(D_3)$ forms are perfect cubes, then $k_3/2h(D_3)$ is the proportion of primes for which $H(p) \equiv 0 \pmod 3$, at least by "Dirichlet density." Actual frequency counts show the proportion to be reassuringly close to $1/6$; (with $k_3 = h(D_3)/3$ in the cases treated here).

The congruence properties $H(p) \equiv 0 \pmod 5$, however, provide too few instances in the range $p < 1000$, to make a frequency count meaningful.

The conditions on p which make $H(p) \equiv 0 \pmod 4$ when $e = -1$, are more provocative. The percentage of such $p(<1000)$ varies from 4 per cent (when $m = 37$) to 12 per cent (when $m = 89$). There seems to be no simple explanation (e.g., in terms of linear or quadratic forms). As a matter of curiosity, when $m = 5$, $H(p) \equiv 0 \pmod 4$ for

$$p = 61, 89, 109, 149, 269, 389, 401, 521, 661, 701, 761, 769, 809, 821, 829;$$

when $m = 37$, this holds for

$$p = 53, 101, 181, 293, 349, 397, 593;$$

TABLE I

Summary of Calculation

Columns 1–5 are explained in Section 2, Columns 6–9 in Section 7, Columns 10–11 in Section 6.

1	2	3	4	5	6	7	8	9	10	11†
m	a	b	e	$h(d)$	m_3	f_3	$h(d_3)$	$h(d_3 f_3^2)$	Q	R or p_4
(Series A: $m \not\equiv 1 \pmod 4$, $c = 1$, $d = 4m$, $d_3 = 4m_3$.)										
2	1	1	−1	1	−6	3	2	6	...	41
3	2	1	+1	1	−1	9	1	6	2	3
6	5	2	+1	1	−2	9	1	6	2	−3
7	8	3	+1	1	−21	3	4	12	−2	7
10	3	1	−1	2	−30	3	4	12	...	157
11	10	3	+1	1	−44	3	4	12	2	11
14	15	4	+1	1	−56	3	4	12	−2	7
15	4	1	+1	2	−5	9	2	12	3*	−5*
19	170	39	+1	1	−57	3	4	24	2	19
22	197	42	+1	1	−66	3	8	24	2	−11
23	24	5	+1	1	−69	3	8	24	−2	23
26	5	1	−1	2	−78	3	4	12	...	37
30	11	2	+1	2	−10	9	2	24	5	−6
31	1,520	273	+1	1	−93	3	4	12	−2	31
34	35	6	+1	2	−102	3	4	12	−2	17
35	6	1	+1	2	−105	3	8	24	5*	−7*
38	37	6	+1	1	−114	3	8	24	2	−19
39	25	4	+1	2	−13	9	2	24	3	−13
42	13	2	+1	2	−14	9	4	24	6	−7
(Series B: $m \equiv 1 \pmod 4$, $c = 2$, $d = m$, $d_3 = m_3$.)										
5	1	1	−1	1	−15	3	2	6	...	61
13	3	1	−1	1	−39	3	4	12	...	29
17	8	2	−1	1	−51	3	2	6	...	13
21	5	1	+1	1	−7	9	1	12	3	−7
29	5	1	−1	1	−87	1 (iii)	6	6	...	13
33	46	8	+1	1	−11	9	1	6	−3	11
37	12	2	−1	1	−111	3	8	24	...	53
41	64	10	−1	1	−123	3	2	6	...	5
53	7	1	−1	1	−159	3	10	30	...	17
57	302	40	+1	1	−19	9	1	12	3	−19
61	39	5	−1	1	−183	3	8	24	...	59
65	16	2	−1	2	−195	3	4	12	...	29
69	25	3	+1	1	−23	1 (i)	3	3	−3	23
73	2,136	250	−1	1	−219	3	4	12	...	37
77	9	1	+1	1	−231	1 (ii)	12	12	7	−11
85	9	1	−1	2	−255	1 (ii)	12	12	...	101
89	1,000	106	−1	1	−267	3	2	6	...	73
93	29	3	+1	1	−31	1	3	3	3	−31
97	11,208	1,138	−1	1	−291	3 (i)	4	12	...	53

* Here $S = 2$. (See Section 6).

† When $e = -1$, Column 11 has the earliest prime p_4 for which $H(p_4) = 4$. (See Section 6).

Table II

Some Special Values of p for Which $n \mid H(p)$

The table gives the minimum odd prime $p(<1000)$ for which $H(p) = n$, (or $H(p) = 2n$, if n is odd and $e = N(\epsilon) = +1$). If no such p occurs, the table lists p_r the earliest prime (<1000) for which $H(p)/n$ (or $H(p)/2n$) gives the minimum quotient r.

m	$n = 8$	16	32	9	27	5	25	7	11
				Series A					
2	137	353	...	269_2	...	79	...	643	199
3*	313	193	...	181	...	71	...	...	...
6*	409	97	...	89	971_2	311	...	743	109
7*	71	751	127	179	271	131	...	197	617_4
10	241	449	...	271	...	19	...	419	131
11*	97	881	449	719	...	409	199	421	...
14*	71	79	...	251	...	29	...	97	...
15*	31	...	...	163	487	61	...	71	...
19*	73	...	...	991	269	31	...	13	397
22*	353	401	641	883	593	271	701	127	131_2
23*	41	47	...	521	...	59	...	631	...
26	641	881	...	...	...	139	...	337_2	...
30*	23	383	...	739	...	439	349	211	...
31*	7	193	...	883_7	...	19	449	13	...
34*	23	911_3	...	163	...	59	...	83	433_4
35*	47	449	223	71	...	89	...	701	...
38*	137	769	...	37	701	431	...	127	...
39*	673	79	...	827	...	151	...	911	857
42*	103	673_7	...	809	431	491	...	433	...
				Series B					
5	89	...	...	919	...	211	...	307	967
13	233	...	...	827	...	59	...	211	109
17	281	...	...	127	...	79	...	...	...
21*	199	337	...	...	...	101	...	433	263
29	233	673	...	971	...	619	...	601_2	461_6
33*	71	47	...	433_2	379	139	...	239	331
37	...	...	...	73_2	...	71	...	167	...
41	769_2	769	...	307	...	199	...	491	593_2
53	929_2	929	449	433_4	...	379	...	113_2	659
57*	487	127	...	197	...	271	...	43	...
61	937	977	...	271	487	59	...	463	...
65	601	...	353	467	431	211	...	...	43
69*	71	239	...	307	...	79	...	97	...
73	857	...	...	107	...	379	...	333_2	67
77*	127	113	...	...	...	101	...	71	...
85	...	...	...	71	...	331	...	139	947
89	809	641	929	631	...	59	...	503	967
93*	463	79	...	379	811_3	251	...	29	947
97	113_2	113	673	107	...	151	...	463	...

(* Denotes values of m for which $N(\epsilon) = 1$).

TABLE III

Quadratic Forms Which are Perfect Cubes

These are the forms (A, B, C) of discriminant $B^2 - 4AC = d_3 f_3^2$ which represent those primes $p(\nmid 6m)$ for which $3 \mid H(p)$, where p is "conveniently" small.

m	$d_3 f_3^2$	A	B	p	$H(p)$	A	B	p	$H(p)$
				Series A					
2	−216	1	0	79	3	2	0	29	6
3	−324	1	0	97	12	2	2	41	6
6	−648	1	0	163	6	2	0	83	12
7	−756	1	0	193	12	7	0	139	6
		2	2	107	6	14	14	17	6
10	−1080	1	0	271	9	2	0	137	6
		5	0	59	3	10	0	37	12
11	−1188	1	0	313	24	11	0	71	6
		2	2	149	6	19	16	19	6
14	−1512	1	0	379	42	2	0	191	24
		7	0	61	6	14	0	41	6
15	−1620	1	0	409	12	5	0	101	6
		2	2	227	12	10	10	43	6
19	−2052	1	0	577	36	19	0	103	6
		2	2	257	6	23	8	23	6
22	−2376	1	0	619	6	2	0	347	12
		11	0	227	12	22	0	331	6
		7	±2	7	6	14	±12	47	6
23	−2484	1	0	877	6	23	0	131	6
		2	2	311	24	25	4	349	6
		5	±4	5	6	10	±6	67	6
26	−2808	1	0	727	3	2	0	353	6
		13	0	67	3	26	0	53	6
30	−3240	1	0	811	30	2	0	503	12
		5	0	167	12	10	0	241	60
		11	±4	11	6	22	±4	37	6
31	−3348	1	0	853	6	27	0	139	6
		2	2	419	30	29	4	29	6
34	−3672	1	0	919	6	2	0	461	6
		17	0	71	24	27	0	61	6
35	−3780	1	0	1009	12	5	0	269	6
		7	0	163	6	27	0	167	12
		2	2	557	6	31	8	31	6
		10	10	97	6	14	14	71	18
38	−4104	1	0	1051	6	2	0	521	6
		19	0	73	24	27	0	179	36
		23	±6	23	6	31	±22	31	6
39	−4212	1	0	1069	12	13	0	337	..
		2	2	587	6	26	26	47	6
		17	±2	17	6	31	±2	31	6
42	−4536	1	0	1303	6	2	0	569	6
		7	0	337	84	14	0	137	6
		13	±12	13	6	26	±12	59	12

TABLE III—*Continued*

m	$d_2 f_2^2$	A	B	p	$H(p)$	A	B	p	$H(p)$
				Series B					
5	−135	1	1	139	3	5	5	47	3
13	−351	1	1	367	3	10	10	79	3
		8	±1	11	3	..	...	...	..
17	−459	1	1	127	9	11	5	11	3
21	−567	1	1	571	6	7	7	109	12
		8	±3	23	6	..	...	...	..
29	−87	1	1	103	3	3	3	41	6
33	−891	1	1	223	6	11	11	23	12
37	−999	1	1	1063	3	16	5	619	3
		2	±1	131	3	4	±3	73	18
		8	±5	89	6	..	...	...	..
41	−1107	1	1	277	12	17	7	71	3
53	−1431	1	1	1447	3	20	13	239	3
		7	±5	7	3	18	±15	23	3
		10	±3	43	3	8	±3	83	3
57	−1539	1	1	397	36	19	19	139	6
		5	±1	5	6	..	...	...	..
61	−1647	1	1	1663	3	22	17	271	9
		18	±3	23	3	8	±7	53	6
		13	±11	13	6	..	...	...	..
65	−1755	1	1	439	3	23	19	23	3
		5	5	89	18	13	13	37	6
69	−23	1	1	101	6	..	...	...	..
73	−1971	1	1	499	3	25	23	79	3
		5	±3	5	6	..	...	...	..
77	−231	1	1	331	6	3	3	89	18
		8	5	233	6	7	7	61	6
85	−255	1	1	271	15	8	1	83	21
		3	3	97	12	5	5	131	3
89	−2403	1	1	601	12	27	27	83	3
93	−31	1	1	47	6	..	...	...	..
97	−2619	1	1	661	12	27	27	31	3
		23	±7	23	3	..	...	...	..

and when $m = 89$, this holds for

$$p = 53, 73, 109, 157, 233, 257, 269, 449, 461, 509, 601, 613, 641, 733, 757, 809, 821, 929, 937, 977.$$

Curiously enough, when $m = 37$ all $p(<1000)$ for which $H(p) \equiv 0 \pmod 4$ satisfy $H(p) = 4$; from Table II, this value of m seems most "resistant to variety" in the values of $H(p)$.

Table IV
Irregular (Odd) Primes < 1000

For values of m in Table I. Primes cf index 2 are marked with (*), unmarked primes are of index 1. (See Section 8.)

m	p	$H(p)$	m	p	$H(p)$
	Series A			Series B	
2	13	2	13	241	2
2	31	1	29	3*	1
3	103	2	29	11	1
6	3	1	33	3	1
6	7	2	33	29	2
10	191	5	33	37	4
10	643	1	37	7	1
15	3	1	37	89	6
15	181	2	37	257	6
19	79	2	41	29*	2
22	43	4	41	53	2
22	73	2	53	5	2
23	7	2	57	59	2
23	733	2	69	5	2
31	157	2	69	17*	2
34	37	2	73	5*	6
34	547	26	73	7	1
35	23	2	73	41	2
38	5	2	85	3	1
39	5	2	89	5*	2
39	7	2	89	7	1
39	37	2	89	13	2
42	3*	1	89	59	5
42	5	2	93	13	2
42	43	2	97	17	2
42	71	2			

It is our hope that additional motivation might be suggested by these data before the next electronic tour de force is attempted.

Department of Mathematics
University of Arizona
Tucson, Arizona, and

Applied Mathematics Division
Argonne National Laboratory
Argonne, Illinois

[1] H. Cohn, "A numerical study of Dedekind's cubic class number formula," *J. Res. Nat. Bur. Standards*, v. 59, 1957, p. 265–271. (A similar composition problem is treated on a computer here.)

2. E. C. Dade, O. Taussky, & H. Zassenhaus, "On the semi-group of ideal classes in an order of an algebraic number field," *Bull. Amer. Math. Soc.*, v. 67, 1961, p. 305–308. (The ideals which divide *f*, normally excluded from class number considerations, are treated theoretically and with the aid of computers.)

3. R. Dedekind, "Über die Anzahl der Idealklassen in reinen kübischer Zahlkörpern," *J. Reine Angew. Math.*, v. 121, 1900, p. 40–123.

4. P. G. L. Dirichlet, "Une propriete des formes quadratiques a determinant positive," *J. Math. Pures Appl.* Ser II, v. 1, 1856, p. 76–79.

5. R. Fueter, *Vorlesungen über die singulären Moduln und die komplexe Multiplikation der elliptischen Funktionen*, v. 1, II Leipzig-Berlin, 1924. (This work gives the classic application of relative class structures concisely.)

6. H. Hasse, "Arithmetische Theorie der kubischen Zahlkörper auf klassenkörpertheoretischer Grundlage," *Math Z.* v. 31, 1929, p. 565–582. (Reference centers primarily on the top line of the table on p. 568.)

7. H. Weber, *Lehrbuch der Algebra*, v. II, III, Braunschweig, 1894, 1908.

8. H. Weber, "Beweiss des Satzes dass jede eigentliche primitive quadratische Form unendlich viele Primzahlen darstellen fähig ist," *Math. Ann*, v. 20, 1882, p. 301–329.

A Very High-Speed Digital Number Sieve

By D. G. Cantor, G. Estrin, A. S. Fraenkel, and R. Turn

1. Introduction. The general sieve problem may be stated as follows [3]. Let $m_1, m_2, \cdots, m_s$ be s positive integers, relatively prime in pairs. Consider the congruences

$$x \equiv a_{ij} (\text{mod } m_i), \qquad i = 1, 2, \cdots, s; j = 1, 2, \cdots, t_i < m_i. \tag{1}$$

For fixed i, the a_{ij} are distinct non-negative integers less than m_i. The problem is to find all integers N between given limits, say

$$A \leqq N < B, \tag{2}$$

such that N is a solution to s of the congruences. (It is, of course, clear that no N can be a solution to more than s of the congruences (1).)

Examples: On the one extreme there is the *Sieve of Eratosthenes* for finding all primes p in the range $A = B^{1/2} \leqq p < B$, where $t_i = m_i - 1$ for all i. (Here m_i are all the primes $< B^{1/2}$.) On the other extreme there is the Chinese remainder type of problem, where $t_i = 1$ for all i, and there is only one solution among $\prod_{i=1}^{s} m_i$ numbers.

In between these two extremes, there is the important *quadratic sieve*, where roughly $t_i = m_i/2$ for all i. It is used in problems involving quadratic residues, Diophantine equations of second degree and other quadratic type problems.

About thirty years ago, Lehmer [1], [2] constructed a novel special-purpose device for sifting. It used the first 30 primes as moduli. Its processing rate was

$$3 \times 10^5 \text{ numbers/min.}$$

General-purpose computers are not very well suited to sifting, and the earlier models could not compete with Lehmer's machine. However, the speed of the more recent machines makes up for their lack of orientation towards the sieve problem insofar as surpassing the performance of Lehmer's machine is concerned. Thus, the rate for a quadratic sieve using the first 30 primes on the IBM 7090 is approximately

$$10^7 \text{ numbers/min.}$$

The present paper describes a special-purpose device, where rates in excess of

$$10^{10} \text{ numbers/min.}$$

can be achieved for quadratic sieves. The device consists of basic digital building blocks from which a suitable sieve is assembled for each problem. Thus, by an appropriate rearrangement of the building blocks, problems with different moduli can be run. It is also shown that if the device contains a certain minimum amount of hardware and is attached to a general-purpose computer, then problems can be run where, roughly speaking, the number of moduli is not limited any more by the

Received June 23, 1961. The preparation of this paper was sponsored, in part, by the Office of Naval Research and the Atomic Energy Commission.

amount of hardware of the special-purpose device, but only by the size of the memory of the general-purpose computer, and the rate is still of the order of 10^{10} numbers/min. We use a so-called "Fixed Plus Variable Structure Computer" organization for realizing the combination between the special- and general-purpose computers [6]. This also enables one to use the digital building blocks of the sieve for building other special-purpose devices which one might want to associate with the general-purpose computer.

2. Binary Set-Up of the Sieve. For solving the system (1) on a digital computer, we consider a matrix M of size $s \times (B - A)$ with entries c_{ik} $(i = 1, 2, \cdots, s;\ k = A, A + 1, \cdots, B - 1)$, defined by

$$c_{ik} = \begin{cases} 1 \text{ if } k \equiv a_{ij} \pmod{m_i} \\ 0 \text{ otherwise} \end{cases} \qquad (j = 1, 2, \cdots, t_i).$$

Then every column, all of whose entries are 1, corresponds to a solution, and conversely.

Example: Find the primes p such that

$$6 \leqq p < 36. \tag{3}$$

The relevant congruences are $x \equiv 1 \pmod 2$, $x \equiv 1, 2 \pmod 3$, $x \equiv 1, 2, 3, 4 \pmod 5$. The matrix M is given by

$m_i \backslash N$	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
3	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1
5	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0

The columns all of whose entries are 1 correspond to the primes in the range (3).

The rows of M are periodic with period m_i. Thus, the first ordered m_i bits of the ith row determine the rest of this row completely, and we call them the *periodic pattern* e_i of m_i.

3. Method of Solution on the Special-Purpose Computer. We now give an informal introduction to the principle of operation of the special-purpose device. It will be observed that the method is based on ideas used in earlier work [1], [2], [3] in this field.

A first approach to the mechanization of a special-purpose sieve would be to build a matrix precisely in the form displayed for the example above with observation posts in every column, detecting coincidences of non-zero bits. Problems which can be solved by such a procedure are limited to those which can fit into the maximum size matrix which can be assembled, i.e., computationally trivial problems.

We order the moduli so that

$$m_1 < m_2 < \cdots < m_s. \tag{4}$$

Since solutions exist only corresponding to columns with non-zero bits, we may eliminate the m_s-row and many of the components required to detect coincidences, by establishing coincidence gates only in those columns where the m_s-row has non-zero entries.

Large problems may be handled by constructing only m_s columns of the matrix and testing for solutions in these columns in parallel. Entering the next batch of m_s numbers turns out to be equivalent to performing prearranged circular shifts in the $s - 1$ rows. This procedure would require a matrix of size $(s - 1) \times m_s$ with coincidence gates established in the columns as prescribed above. It is possible to use such a matrix to define potential solutions even when it is only feasible to mechanize $l < s - 1$ rows of the matrix, and then have a general-purpose computer complete the test for solution.

The range of problems which may be handled is increased when it is recognized that the periodic pattern e_i associated with the ith row completely determines the rest of the row. In the following we give an algorithm which defines a procedure requiring only m_i elements in the ith row, giving up only the regularity of the coincident gate connections. The special-purpose computer consists of basic digital building blocks or modules which are assembled into a matrix consisting of $s - 1$ shifting registers, the ith of length m_i, and initially containing the periodic pattern e_i $(i = 1, 2, \cdots, s - 1)$. Observation posts are placed at certain positions in the matrix which sift out the solutions to (1) among the first m_s numbers.* Next, a circular shift is performed in each register, which is equivalent to bringing in the next m_s numbers to be sifted. This is followed by the observation posts sifting out the solutions among this new batch of numbers. This process of sifting followed by shifting is continued until all the numbers are processed.

4. The Algorithm. We divide the numbers N in the range (2) into sets S_n defined by†

$$S_n = \{N : N < B, N = A + nm_s + k; 0 \leqq k < m_s\}, \quad n = 0, 1, \cdots, \left[\frac{B - A - 1}{m_s}\right]. \tag{5}$$

Thus, each set (except possibly the last) contains m_s numbers.

Let

$$m_s = q_i m_i + r_i, \qquad 0 < r_i < m_i \quad (i = 1, 2, \cdots, s - 1). \tag{6}$$

With each set S_n we associate a matrix M_n of size $(s - 1) \times m_s$ with entries $c_{ij}(n)$ $(i = 1, 2, \cdots, s - 1; j = 0, 1, \cdots, m_s - 1)$, defined recursively by

$$c_{ij}(0) = \begin{cases} 1 \text{ if } 0 \leqq j < m_i \quad \text{and} \quad A + j \equiv a_{ij}, \cdots, a_{i,t_i} \pmod{m_i} \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

$$c_{ij}(n) = \begin{cases} c_{i,j+r_i}(n - 1) & \text{if} \quad 0 \leqq j + r_i < m_i \\ c_{i,j+r_i-m_i}(n - 1) & \text{if} \quad 0 \leqq j < m_i \quad \text{and} \quad j + r_i \geqq m_i \\ 0 & \text{if} \quad m_i \leqq j < m_s . \end{cases} \tag{8}$$

* The positioning of the observation posts is determined by m_s and its residues in such a way that a register of length m_s is not required.

† $[x]$ stands for the largest integer $\leqq x$.

Equation (8) can be written in the form

$$c_{ij}(n) = \begin{cases} c_{id}(n-1) \quad \text{where} \quad m_i > d \equiv j + r_i \pmod{m_i}, & \text{if} \quad 0 \leqq j < m_i \\ 0 \quad \text{if} \quad m_i \leqq j < m_s . \end{cases}$$

Hence (7) and (8) are equivalent to

$$c_{ij}(n) = \begin{cases} 1 \quad \text{if} \quad 0 \leqq j < m_i \quad \text{and} \quad A + j \equiv a_{i,v_i} - nr_i \pmod{m_i} \\ 0 \quad \text{otherwise} \qquad\qquad (v_i = 1, 2, \cdots, t_i). \end{cases} \tag{9}$$

If $N \epsilon S_n$ is a solution to the system (1), then by (5),

$$A + k \equiv a_{s,v_s} \pmod{m_s} \quad (0 \leqq k < m_s ; v_s = 1, 2, \cdots, t_s) \tag{10}$$

for all n.
By (5) and (6) we have also

$$A + k \equiv a_{i,v_i} - nr_i \pmod{m_i} \quad (v_i = 1, 2, \cdots, t_i ; i = 1, 2, \cdots, s-1). \tag{11}$$

Hence, if we let

$$k = w_i m_i + u_i , \qquad 0 \leqq u_i < m_i \qquad (i = 1, 2, \cdots, s-1), \tag{12}$$

then by (11),

$$A + u_i \equiv a_{i,v_i} - nr_i \pmod{m_i} \qquad (v_i = 1, 2, \cdots, t_i ; i = 1, 2, \cdots, s-1),$$

so that

$$c_{i,u_i}(n) = 1 \tag{13}$$

for $i = 1, \cdots, s-1$ by (9).

Also the converse holds. That is to say, if (13) holds for $i = 1, \cdots, s-1$ (where u_i is given by (12) and k by (10)), then

$$N = A + nm_s + k \tag{14}$$

is a solution to the system (1). This is the basis of the algorithm. We list the t_s solutions $k_1, \cdots, k_{t_s}$ of (10), and for each of them its corresponding $s - 1$ values u_i. Then the numbers $N = A + nm_s + k_{v_s}$ $(v_s = 1, \cdots, t_s)$ for which (13) holds for $i = 1, \cdots, s-1$, are solutions to (1), and these are all the solutions.

Example: Find all solutions in the range

$$-15 \leqq N < 25$$

to the system of congruences

$$\begin{aligned} x &\equiv 1 && \pmod{2} \\ x &\equiv 1, 2 && \pmod{3} \\ x &\equiv 2, 3, 4 && \pmod{5} \\ x &\equiv 0, 1, 2, 5 && \pmod{7} \\ x &\equiv 0, 1, 8, 9 && \pmod{11}. \end{aligned}$$

Reference is made to Table I. The matrix M_0 is constructed according to (7) (omitting all strings of zeros). Equation (8) implies that the ith row of matrix M_n is obtained from the ith row of M_{n-1} by means of circularly left shifting the

TABLE I

The Matrices for the Problem

N	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5
i j	0	1	2	3	4	5	6	7	8	9	10
1	**1**	0									
2	0	1	**1**						M_0		
3	0	0	**1**	1	1						
4	0	1	**1**	1	0	0	1				

N	−4	−3	−2	−1	0	1	2	3	4	5	6
i j	0	1	2	3	4	5	6	7	8	9	10
1	0	1									
2	1	0	1						M_1		
3	0	1	1	1	0						
4	0	0	1	0	1	1	1				

N	7	8	9	10	11	12	13	14	15	16	17
i j	0	1	2	3	4	5	6	7	8	9	10
1	1	0									
2	1	1	0						M_2		
3	1	1	1	0	0						
4	1	1	1	0	0	1	0				

N	18	19	20	21	22	23	24
i j	0	1	2	3	4	5	6
1	0	**1**					
2	0	**1**	**1**			M_3	
3	**1**	**1**	0	0	1		
4	0	**1**	0	1	1	**1**	0

first m_i bits by r_i positions (or by circularly right shifting them by $m_i - r_i$ positions). In the present case, $r_1 = 1, r_2 = 2, r_3 = 1, r_4 = 4$, so that in passing from one matrix to the next, the first row is shifted left circularly by 1 position, the second by 2, the third by 1 and the fourth by 4 positions. This is the way M_1, M_2 and M_3 are obtained.

The values k_1, k_2, k_3, k_4 of Table II are computed by (10), and the corresponding values of u_i by (12). Table II defines a pattern of observation stations which sift out the solutions in each matrix; the entry u_i represents the column coordinate corresponding to the row coordinate i, at which an observation station exists. In M_0, the observation pattern $u_i = 0, 2, 2, 2$ indicates a solution, since all matrix positions corresponding to that pattern are filled by 1's. They appear in bold type in Table I. The corresponding value of k is $k_4 = 2$. The solution is, therefore, $N = -15 + 0 \times 11 + 2 = -13$ by (14). The only two other solutions

TABLE II
Observation Posts for the Problem

i	m_i	U_i			
		$k_1 = 4$	$k_2 = 5$	$k_3 = 1$	$k_4 = 2$
1	2	0	1	1	0
2	3	1	2	1	2
3	5	4	0	1	2
4	7	4	5	1	2

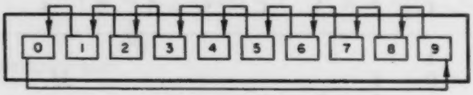


FIG. 1.—Conventional Shift Register.

are found in M_3, for $k_3 = 1$ and $k_2 = 5$, also indicated by bold type in Table I. They are $N = -15 + 3 \times 11 + 1 = 19$ and $N = -15 + 3 \times 11 + 5 = 23$.

5. The Special-Purpose Computer. The special-purpose device should be flexible enough to allow usage of different moduli. Therefore the basic building blocks of the device are modules consisting of memory elements and gates which will be assembled into the appropriate-size shifting registers and sets of and-gates for each problem. The and-gates are the realization of appropriate combinations of the observation posts, and test for coincidence of 1's.

The example of the previous section suggests the construction of the device. Its central part consists of shift-registers $R_1, \cdots, R_{s-1}$, the ith of length m_i, which will store the matrices M_n (without the trailing strings of 0's). Register R_i will shift circularly left by r_i positions. It is important to note, however, that this can be effected in one shift time, rather than in r_i shift times, and further, that the wiring can be so arranged that any transmitting memory element is adjacent to its receiver. In order to do this, we rename the memory elements in R_i so that element number 0 is at the left, followed by element number r_i, followed by number $2r_i \pmod{m_i}$, by $3r_i \pmod{m_i}$, etc. Then each transmitting element is adjacent to its receiver, and every element will appear exactly once. For‡ $(m_i, m_s) = 1$, so that also $(m_i, r_i) = 1$ $(i = 1, \cdots, s - 1)$. Hence r_i generates the additive cyclic group of non-negative integers $\pmod{m_i}$ and every element appears exactly once in the register.

Example: Suppose that for a certain sieve problem $m_s = 23$, and $m_i = 10$ for some $i < s$. Then R_i has to shift circularly left by $r_i = 3$ positions. On a conventional shift-register, three shifts of the type indicated in Figure 1 would have to be performed. The same result can be obtained in one shift time by the specially wired-up register of Figure 2. However, the long wiring has an undesirable effect on the speed of the system. Renaming the memory elements as in Figure 3, the three shifts can be done in one shift time with conventional wiring.

‡ (a, b) stands for the greatest common divisor of a and b.

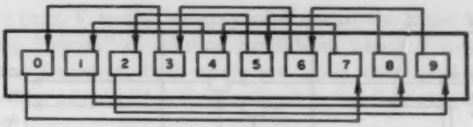


Fig. 2.—Specially Wired-Up Register for Performing a Shift of Three Positions in One Shift Time.

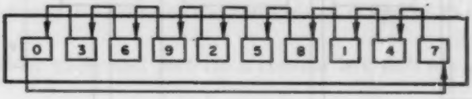


Fig. 3.—Final Form of the Register.

The second part of the special-purpose device consists of t_s sets of and-gates, one set for each value of k which is a solution to (10), "anding" together positions c_{ij} in the registers, as defined in the previous section. Also a counting register R with capacity $> (B - A)/m_s$ and a shift control are required. (The shift control does not normally have to be reconstructed for every problem.) The register R contains the value n of the matrix M_n currently being sifted.

The sifting process consists of the following steps.

1. Clear R.
2. Load $R_1, \cdots, R_{s-1}$ with M_0, whose entries are defined by (7).
3. Record n and record the values k for which coincidence is obtained, i.e., the values k associated with the sets of and-gates which are excited, if any. The corresponding solutions are given by (14).
4. Advance R by unity and perform a circular shift in each shift-register according to the above scheme. This effectively reloads the registers with the next batch of m_s numbers to be sifted.
5. Terminate process if $n > (B - A)/m_s$. Otherwise go back to 3.

It is thus seen that in this process m_s numbers are processed per shift time.

6. The Sieve in a Fixed Plus Variable Structure Computer. It was remarked by Lehmer that special-purpose equipment attached to the arithmetic unit of a fast computer can speed up computation of permutation problems [4], and of other problems [5]. More generally, we consider a so-called "Fixed Plus Variable Structure Computer" (to be designated by $(F + V)$ computer), which consists of a conventional digital computer (the fixed part to be denoted by F), and a set of modules (the variable part to be denoted by V). Many problems contain a part which can be solved on a special-purpose computer in a much more efficient way than on a general-purpose computer. For such a problem, the modules are assembled into a suitable special-purpose device which handles this part. The rest of the problem is handled by F. A supervisory control coordinates the operation of the two computers. However, the special-purpose configuration is not retained permanently, but may be reorganized into other configurations for other problems. For a more detailed description of the concept of the $(F + V)$ computer, the reader is referred to the literature [6].

The special-purpose device described above has a limited amount of hardware.

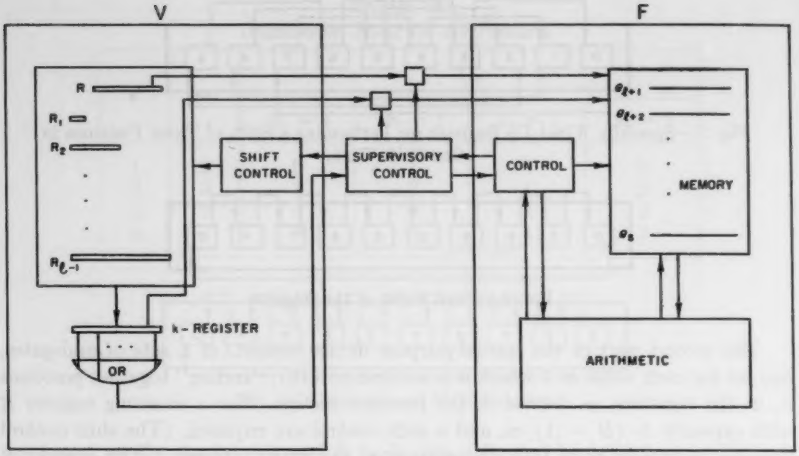


FIG. 4.—$(F + V)$ Computer Organization for the Sieve Problem.

For certain problems it may be desirable to use more moduli than can be mechanized with the available hardware. In order to handle such problems, we imbed the special-purpose device in an $(F + V)$ computer. This allows, as will be shown subsequently, handling problems in which the number of moduli is limited only by the number of periodic patterns e_i of m_i that can be stored in the memory of F, provided that the hardware of V is sufficient to mechanize the first l of the s moduli. The parameter l depends on the relative number of 1's and 0's in the periodic patterns of the sieve, and on the relative speeds of F and V. The use of $F + V$ also allows using the modules for building special-purpose devices other than the sieve, and attaching them to F.

Figure 4 shows the organization of the $(F + V)$ computer for the sieve problem by means of a block diagram. The V-part, which acts as the special-purpose device, mechanizes the first l moduli. The k-register records the values of k (the solutions of (10)) corresponding to coincidence. The periodic patterns $e_{l+1}, \cdots, e_s$ of $m_{l+1}, \cdots, m_s$ are stored in the memory of F. Numbers will be sifted in V, and when coincidence occurs, the contents of R and of the k-register are transferred to F, where so-called ***solution candidates*** N of the form (14) are formed, one for each value of k. Then divisions of the form

$$N = \sigma_i m_i + \rho_i, \qquad 0 \leqq \rho_i < m_i \qquad (i = l + 1, \cdots, s) \tag{15}$$

are performed in F. The residue ρ_i determines uniquely the position of the bit of the periodic pattern e_i of m_i corresponding to N. The number N is a solution if and only if these bits are 1 for $i = l + 1, \cdots, s$. Thus, the $(F + V)$ computer is so organized that V will do the high-speed sifting, and F will do divisions. l will be chosen so that the average time per coincidence in V is at least as large as the average division time per coincidence in F. Then V would normally do its divisions, until such time when coincidence is obtained in V. At such time, V is interrupted and the transfers from V to the memory of F occur, the latter acting as a buffer, capable of storing "bursts" of solution candidates which V might produce occasionally. After the transfers, both parts are again decoupled and assume their respective tasks. The program for V is outlined in the following six steps.

1. Clear R.
2. Load $R_1, \cdots, R_{l-1}$ with M_0.
3. Check for coincidence. If none is obtained, go to 5. Otherwise continue with 4.
4. Interrupt F and V. Transfer the contents of R and of the k-registers to the memory of F.
5. Advance R by unity and perform a circular left shift of r_i places in R_i $(i = 1, \cdots, l-1)$.
6. Terminate process if $n > (B - A)/m_l$. Otherwise go back to 3.

The program for F is simply to produce solution candidates N of the form (14), and to perform divisions of the type (15) for each of them until the first 0-bit is encountered. If none is encountered, N is recorded as a solution.

7. Speed and Hardware. The speed and hardware requirements will now be discussed in terms of an example, for which we choose a quadratic sieve problem where the moduli are the first s primes. The first column of Table III contains values of the independent variable l, the number of moduli mechanized in the special-purpose device. The table displays the speed and hardware requirements for such a sieve as a function of l. The second column contains the lth prime. Let t be the total time required for performing the coincidence test and the subsequent circular shifts in the registers. Using the register organization described in Section 5, the circular shifting amounts to a left shift of one position in each register. We assume transistorized circuitry, for which

$$t = 0.2\ \mu\text{ sec}$$

is chosen (speeds approximating those of the IBM 7090). Thus, m_l numbers are processed in this time if no coincidence occurs. (If the registers are of the double-rank type, both ranks will be equipped with sets of and-gates, and $t = 0.2\ \mu$ sec is the time for a coincidence check and for transferring one rank into the other. Thus also in this case m_l numbers are processed in $0.2\ \mu$ sec.)

We consider first the case $s = l$, that is, we use only a special-purpose computer without a conventional general-purpose computer. Assuming the solutions to be sparse, so that we may neglect the time of recording them, the rate of the sieve is

$$v = \frac{6 \times 10^7 \times m_l}{t} = 3 \times 10^8 \times m_l \text{ numbers/min.}$$

These values are displayed in the third column.

Since the sieve is quadratic, the probability of any randomly selected bit in the periodic pattern e_i to be 1 is about 0.5. Hence, on the average, one coincidence is obtained per 2^l numbers sifted, or every

$$\tau = \frac{6 \times 10^7 \times 2^l}{v} = \frac{2^l}{5m_l}\ \mu\text{ sec.}$$

These values appear in the fourth column.

If $s > l$, divisions have to be performed in F. The probability that exactly i divisions suffice to decide whether any solution candidate N has to be rejected or accepted is $(\frac{1}{2})^i$ $(1 \leqq i \leqq s - l)$. Hence the expectation of the number of divisions for each N is given by

$$\nu = \sum_{i=1}^{s-l} i/2^i = 2 - (s - l + 2)/2^{s-l}. \tag{16}$$

Thus the average number of divisions for each solution candidate approaches 2 asymptotically from below. Assuming the IBM 7090 as the fixed machine F, this division subroutine takes about 200 μ sec for two divisions. Also, preliminary studies of the mode of transfer from V to the 7090 indicate that the transfer of the contents of R and of the k-register requires no more than 7 μ sec. (See appendix.) That is, this is the maximum time during which V is idle. F is interrupted only insofar as it requires memory access during this time. Actually, V could already resume its operation after the transfer of n from the R-register. It would have to wait additional time only if a new solution candidate is formed before the current contents of the k-register has been stored away, which is a rare event. However, in our computation of the overall speed of the sieve we assumed that V is interrupted for 7 μ sec. during each transfer.

Thus, the average overall rate of the sieve is given by

$$w = \frac{v}{1 + 7/\tau} \text{ numbers/min.}$$

for $\tau \geqq 200$ μ sec. If $\tau < 200$ μ sec, V will have to wait for F, and the average overall rate for this case is

$$w = \frac{\tau}{200} \frac{v}{1 + 7/\tau} \text{ numbers/min.}$$

Thus, the operation of the sieve becomes rapidly more and more inefficient as τ decreases below the critical value of 200 μ sec. The values of w appear in the fifth column of Table III. The last column displays the required number $h = \sum_{i=1}^{l-1} m_i$ of memory elements for the special-purpose device. (This number has to be doubled if the registers are of the double-rank type.)

The lower bound for l in Table III was chosen to be 9 because for $l = 8$ the rate would already be less than can be achieved with conventional present-day computers. The upper bound was chosen by setting arbitrarily a hardware constraint of 1500 memory elements.

The lowest value of τ for which $\tau \geqq 200$ μ sec is $\tau = 247.3$ μ sec. Thus V should contain at least 15 registers consisting of 328 memory elements. Figure 5 displays the overall rate as a function of required memory elements. Two simple conclusions can be drawn from the monotonicity of w as displayed in Figure 5. First, l should be chosen as large as possible. That is to say, as much hardware as available should be thrown in to build the sieve; even so the cooperation between F and V becomes less efficient as l increases beyond the critical value of 16, in the sense that F becomes more idle. Secondly, the use of slower memory elements is indicated if a larger number of them is available, hence the possibility of using magnetic core registers.

TABLE III

Speed and Hardware as a Function of l For a Quadratic Sieve

l—No. of Moduli Implemented in Special-Purpose Device	m_l—the lth Modulus	v—Numbers/Min. Rate of Sieve if $s = l$	τ Average Time Per Coincidence	w—Numbers/Min. Rate of Sieve if $s > l$	h—Number of Memory Elements in Special-Purpose Device
9	23	6.9×10^9	4.5 μ sec	5.70×10^7	77
10	29	8.7×10^9	7.1 μ sec	1.56×10^8	100
11	31	9.3×10^9	13.2 μ sec	4.03×10^8	129
12	37	1.11×10^{10}	22.1 μ sec	9.28×10^8	160
13	41	1.23×10^{10}	39.9 μ sec	2.09×10^9	197
14	43	1.29×10^{10}	76.2 μ sec	4.49×10^9	238
15	47	1.41×10^{10}	139.4 μ sec	9.34×10^9	281
16	53	1.59×10^{10}	247.3 μ sec	1.54×10^{10}	328
17	59	1.77×10^{10}	444.3 μ sec	1.73×10^{10}	381
18	61	1.83×10^{10}	859 μ sec	1.81×10^{10}	440
19	67	2.01×10^{10}	1.6 m sec	2.01×10^{10}	501
20	71	2.13×10^{10}	2.9 m sec	2.13×10^{10}	568
21	73	2.19×10^{10}	5.7 m sec	2.19×10^{10}	639
22	79	2.37×10^{10}	10.6 m sec	2.37×10^{10}	712
23	83	2.49×10^{10}	20.2 m sec	2.49×10^{10}	791
24	89	2.67×10^{10}	37.7 m sec	2.67×10^{10}	874
25	97	2.91×10^{10}	69.2 m sec	2.91×10^{10}	963
26	101	3.03×10^{10}	132.9 m sec	3.03×10^{10}	1060
27	103	3.09×10^{10}	260.6 m sec	3.09×10^{10}	1161
28	107	3.21×10^{10}	501.7 m sec	3.21×10^{10}	1264
29	109	3.27×10^{10}	985.1 m sec	3.27×10^{10}	1371
30	113	3.39×10^{10}	1900 m sec	3.39×10^{10}	1480

By (16), the average number of divisions per solution candidate in F is less than 2, whatever the number of periodic patterns that are stored in F. Therefore, the rate w of a quadratic sieve is practically independent of s, and the number of the moduli is limited only by the number of periods that can be stored in F. A similar remark applies also for sieves that are "less than quadratic," i.e., where the number of 1's in e_i is $< m_i/2$. For these types of sieves there are even less divisions to be performed, and a higher overall rate is obtained. For sieves that are "more than quadratic," and in particular for those which approach the type of sieve of Eratosthenes, more than two divisions are required on the average, and a higher critical value of l is obtained.

Variations of the above described method which result in even higher speeds (and therefore involve higher critical values of l) are clearly possible. For example, two moduli may be combined in V, say m_l and m_{l-1}, by initially solving the two congruences involving m_l and m_{l-1} manually or on F. Then $m_l m_{l-1}$ numbers can be processed per shift time, for which $t_l t_{l-1}$ sets of and-gates are required. Registers m_l and m_{l-1} do not have to be built of course. As another example, we might mechanize the moduli $m_1, m_2, \cdots, m_{l-1}, m_s$ in V, rather than $m_1, m_2, \cdots, m_{l-1}, m_l$, so that m_s numbers instead of only m_l are processed per shift time.

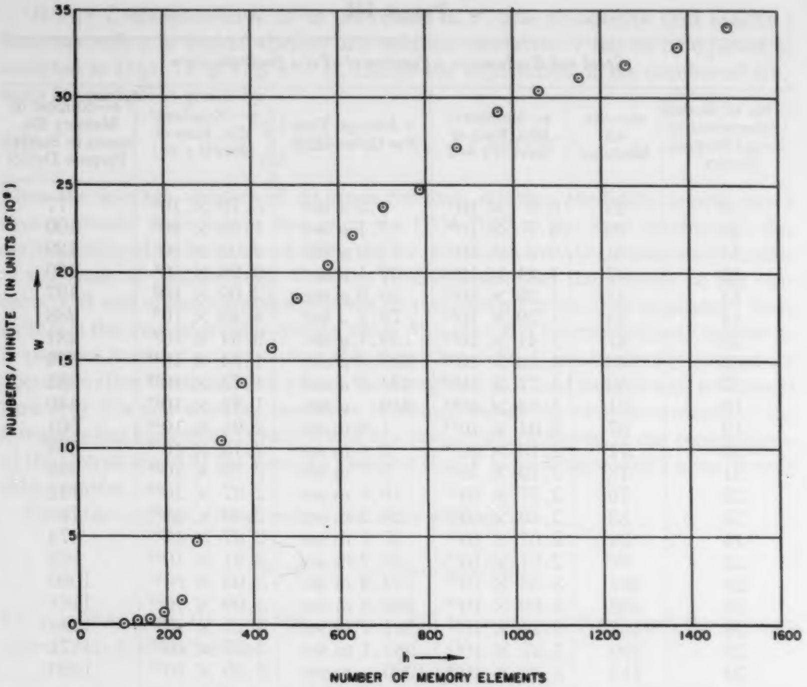


FIG. 5.—Rate of Sieve as a Function of the Number of Memory Elements.

Similarly, two moduli may be combined in F. For example, combining m_{l+1} with m_{l+2} and m_{l+3} with m_{l+4} (which increases storage requirements in F), and using as divisors the moduli $m_{l+1}m_{l+2}$, $m_{l+3}m_{l+4}$, m_{l+5}, $\cdots$, m_s, the average number of divisions that have to be performed approaches 11/8 asymptotically from below. Thus, the effect of combining moduli in F is to lower the critical value of τ. Such a procedure would therefore be used when the available hardware in V is smaller than required for keeping up with the speed of F implied by an average of two divisions per solution candidate.

The high speeds which can be achieved by our method suggest its applicability for conversion of numbers from the modular number representation [7] to the conventional polyadic representation. Since this problem is of the Chinese remainder type, it seems possible to include in the sieve special solution hunting properties.

8. Conclusion. A method has been presented to sift numbers satisfying a set of linear congruences from among a large set of numbers. The important properties of the resulting special-purpose device are that a relatively large set of numbers is processed essentially within the time required for performing a shift of one position in an ordinary shift-register, and that no memory references are necessary. This leads to an overall speed gain of about three orders of magnitude over modern present-day computers such as the IBM 7090. By combining the device with a general-purpose computer, the size of problems that can be run is greatly increased with almost no decrease in speed.

Appendix

The Division Subroutine. For the purposes of this subroutine, written for the IBM 7090, we restrict the size of N in (2) to a number representable by 72 binary bits.

The first 36 of these are called HW, and the last 36 are called LW. The core of the subroutine consists of the following sequence, where it is assumed that the accumulator is cleared at the beginning. Every bit of the periodic patterns e_i is stored in a separate word, denoted by WM, and the corresponding period is stored in M.

```
LDQ   (Load the MQ)                       HW
DVP   (Divide)                            M
LDQ   (Load the MQ)                       LW
DVP   (Divide)                            M
PAC   (Place complement of address        0, 4
      in index register)
CLA   (Clear add)                         WM, 4
TMI   (Transfer on minus)                 OUT
```

This sequence requires 36 cycles. For a quadratic sieve, the sequence has to be performed twice on the average for each solution candidate. Another 20 cycles are required for performing the multiplication and addition implied by (14) and bookkeeping. One cycle takes 2.18 μ sec. Thus the subroutine requires about 200 μ sec.

Transfers from V to F. A preliminary study of the$(F + V)$ organization based on the IBM 7090 as F indicates that transfers from V to F can be effected in the manner of a data channel. Such a channel has a "Channel Address Counter" (CAC), from which addresses are transferred to the "Memory Address Register."

Suppose that the memory region bounded by addresses K and $K + M$ is allocated for storing the value n contained in R, and L to $L + M$ for storing the contents of the k-register. We assume, for simplicity, that registers R and k do not exceed 36 bits. In its normal form, the CAC contains an address of the form $K + i$ $(0 \leqq i \leqq M)$. Three flip-flops $FF1$, $FF2$, $FF3$ are contained in SC. $FF1$ records whether F or V was the last user of the buffer region of the memory. $FF2$ and $FF3$ define "full" and "empty" conditions of the buffer.

We adopt the following operating rules.

1. When V wants to store into the memory, the CAC is advanced by 1 if $FF1 = 1$, and remains unchanged if $FF1 = 0$. Then n is stored at the address currently held in CAC, say $K + i$. Next CAC is changed to $L + i$, and the contents of the k-register are stored. Then CAC is set back to $K + i$. After execution of these stores, $FF1$ is set to 1.

2. When F wants to fetch a pair of new values from the memory, the CAC is decreased by 1 if $FF1 = 0$, and is left unchanged if $FF1 = 1$. The address (con-

tents of CAC) is forced into the F Memory Address Register as a consequence of recognition of a special instruction in F by the Supervisory Control. Both the value n and the corresponding contents of the k-register are then fetched by the previously described $K - L$ interchange, and CAC is set back to $K + i$. At the end of the fetching operations, $FF1$ is set to 0.

Thus F always handles first the latest information brought in from V. If at any time CAC holds the address $K + M$, and if $FF1 = 1$, then $FF2$ is set, which prevents V from storing into the memory. (Of course for sufficiently large l, such an occurrence is very rare.) $FF2$ is reset by the resetting signal of $FF1$. If at any time CAC holds the address K, and if $FF1 = 0$, then $FF3$ is set, which prevents F from fetching. $FF3$ is reset by the setting signal of $FF1$.

Preliminary studies of this mode of transfer indicate that transfer of the first word takes at most two cycles, and the second takes one cycle. Thus the transfer of n and the contents of the k-register from V to F requires approximately 7 μ sec.

Department of Mathematics
Princeton University
Princeton, New Jersey

Department of Engineering
University of California
Los Angeles 24, California

Department of Mathematics
University of Oregon
Eugene, Oregon

Department of Engineering
University of California
Los Angeles 24, California

1. D. H. Lehmer, "A photo-electric number sieve," *Amer. Math. Monthly*, v. 40, 1933, p. 401–406.

2. D. H. Lehmer, "A machine for combining sets of linear congruences," *Math. Ann.* v. 109, 1934, p. 661–667.

3. D. H. Lehmer, "The sieve problem for all-purpose computers," *MTAC*, v. 7, 1953, p. 6–14.

4. D. H. Lehmer, "Teaching combinatorial tricks to a computer," *Proceedings of Symposia in Applied Mathematics*, v. X, American Mathematical Society, Providence, R.I., 1960, p. 179–193.

5. D. H. Lehmer, private communication, June 1960.

6. G. Estrin, "Organization of computer systems—the fixed plus variable structure computer," *Proceedings of the Western Joint Computer Conference*, May 1960, p. 33–37.

7. A. Svoboda, "Rational numerical system of residual classes," *Stroje Na Zpracovani Informaci*. (Czechoslovakia), v. 5, 1957, p. 9–47.

On Finite Difference Methods of Solution of the Transport Equation

By R. P. Pearce and A. R. Mitchell

1. Introduction. In recent years several difference schemes have been proposed for solving the transport equation

$$\frac{\partial u}{\partial t} + V(x, t, u)\frac{\partial u}{\partial x} = F(x, t, u) \tag{1}$$

in one form or another, where V is the velocity of propagation of a profile given initially along the x-axis. Most of these schemes can be found in Richtmyer [1] and, generally speaking, they are chosen primarily from the point of view of stability.

An equation of the type (1) has a single family of characteristics in the (x, t) plane and in any step-by-step method of solution it is essential from the point of view of accuracy that the characteristics be followed as closely as possible. It is proposed to examine existing difference schemes from this standpoint and to derive new formulas of greater accuracy. For the purposes of this paper, it is sufficient to consider the simplified version of (1)

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} = 0, \tag{2}$$

where V is constant, from which it follows that the given profile at $t = 0$ is propagated without change of shape in the direction of the x-axis with velocity V. If a difference scheme fails to give an accurate solution of (2), it is pointless to consider it as a means of solving more complicated forms of (1), in particular, forms which incorporate variable velocity of propagation and source or sink terms. On the other hand, it is realized that schemes which successfully solve (2) may not give comparable accuracy when used to solve (1). In the case of (1), the characteristics are curved and can only be determined by integration of the equation $\frac{dx}{dt} = V(x, t, u)$. In addition, the equation $\frac{du}{dt} = F(x, t, u)$ has to be solved. These computations, however, involve only numerical integration, a process which can be made as accurate as required in most problems.

2. Stable Finite Difference Schemes Now in Use. Existing stable difference schemes will now be discussed with reference to equation (2). The characteristics of the latter are straight lines inclined to the t-axis at an angle

$$\theta = \tan^{-1} V. \tag{3}$$

In these schemes, the parameter p is introduced where $p = \frac{V\Delta t}{\Delta x}$, and Δx and Δt are the respective mesh lengths in the x and t directions.

Received January 10, 1961.

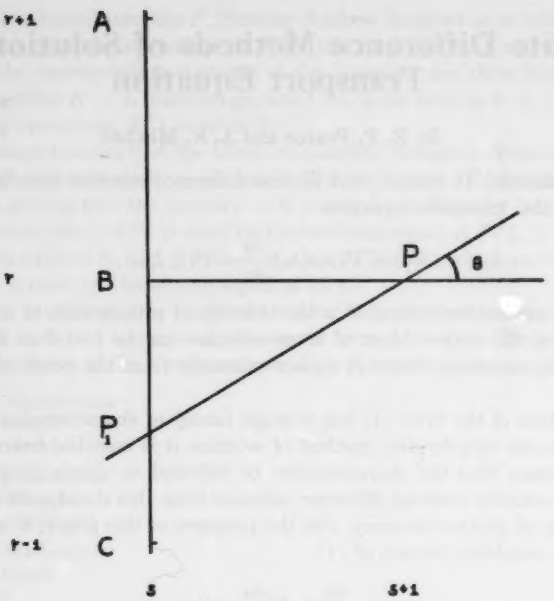


FIG. 1

Difference System I (*Friedrichs* [2]). This is given by

$$u_{r,s+1} = \tfrac{1}{2}(1 - p)u_{r+1,s} + \tfrac{1}{2}(1 + p)u_{r-1,s} \tag{4}$$

where $x = r\Delta x$ and $t = s\Delta t$. This system can be obtained by replacing $\dfrac{\partial u}{\partial x}$ and $\dfrac{\partial u}{\partial t}$ at the node (r, s) by $\dfrac{1}{2\Delta x}(u_{r+1,s} - u_{r-1,s})$ and $\dfrac{1}{\Delta t}(u_{r,s+1} - u_{r,s})$ respectively, then substituting $\frac{1}{2}(u_{r+1,s} + u_{r-1,s})$ for $u_{r,s}$. Another and more satisfactory way of deriving (4) is now proposed. In Figure 1, the characteristic through P cuts AC in P_1 where $BP_1 = p\Delta x$, and it follows that $u_P = u_{P_1}$. Since P_1 is not a mesh point, the value of u at P_1 may be obtained by linear interpolation between A and C, and so (4) is obtained. In addition, since the coefficients on the right-hand side of (4) have sum unity, the solution computed by (4) is bounded if both coefficients are positive which leads immediately to the condition $| p | \leqq 1$ for stability (Richtmyer [1], p. 43).

Difference System II (*Carlson* [3]). This system is given by

$$\text{(5a)} \quad u_{r,s+1} = (1 - p)u_{r,s} + pu_{r-1,s} \qquad (0 \leqq p \leqq 1)$$

$$\text{(5b)} \quad u_{r,s+1} = \frac{1}{1 + p}\, u_{r,s} + \frac{p}{1 + p}\, u_{r-1,s+1} \qquad (p > 1)$$

and two similar formulas if $p < 0$.

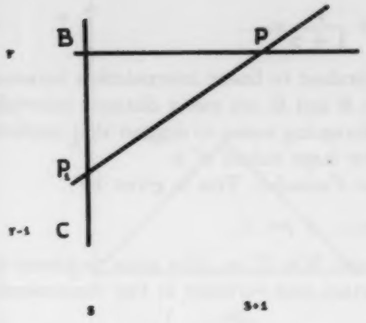


FIG. 2(a)

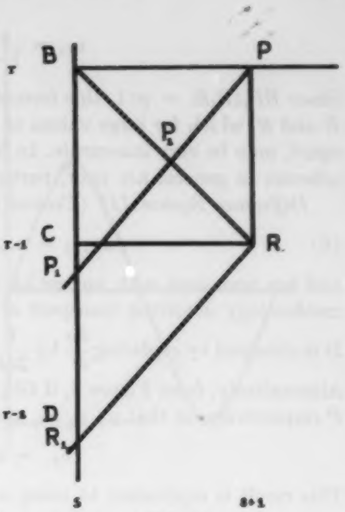


FIG. 2(b)

In Figure 2, PP_1 is the characteristic through P. In this scheme, three points only are used, the choice of points depending on the position of P_1. If $0 \leqq p \leqq 1$, P_1 lies between B and C and the formula used is (5a), whereas if $p > 1$, P_1 lies outside BC and the formula used is (5b). It is presumed that these formulas were obtained originally by replacing $\frac{\partial u}{\partial t}$ by $\frac{1}{\Delta t}(u_{r,s+1} - u_{r,s})$ and $\frac{\partial u}{\partial x}$ by $\frac{1}{\Delta x}(u_{r,s} - u_{r-1,s})$ or $\frac{1}{\Delta x}(u_{r,s+1} - u_{r-1,s+1})$ for $0 \leqq p \leqq 1$ and $p > 1$ respectively.

When P_1 lies between B and C (Figure 2a) it follows that $BP_1{:}P_1C = p{:}1 - p$. Thus, on using linear interpolation between B and C together with the result $u_P = u_{P_1}$, formula (5a) is obtained. When P_1 lies beyond C (Figure 2b) it can be shown that $BP_2{:}P_2R = p{:}1$, and so using linear interpolation between B and R together with $u_P = u_{P_2}$, formula (5b) is obtained. The solution computed by (5) is bounded for all p, as the right-hand sides of both (5a) and (5b) sum to unity and have positive coefficients.

As Carlson's scheme has been used extensively to solve problems involving the transport equation [1], [4], it is worth studying in some detail with a view to determining its probable accuracy. If $0 \leqq p \leqq 1$, formula (5a) is as accurate as the linear interpolation of u between B and C. As these are neighboring mesh points on $t = s\Delta t$, the line of most recently computed values of u, it is to be expected that (5a) will give reasonably accurate values of u. Certainly (5a) will be superior to scheme (4) proposed by Friedrichs since the latter uses linear interpolation of u between A and C, mesh points two distance intervals apart. If $p > 1$, however, a much less satisfactory state of affairs exists. In Figure 2b, RR_1 is the characteristic through R, and theoretically $u_R = u_{R_1}$. Similarly, $u_P = u_{P_1}$ and (5b) becomes

$$u_{P_1} = \frac{1}{1+p} u_B + \frac{p}{1+p} u_{R_1}.$$

Since $BP_1 : P_1R_1 = p:1$, this formula is equivalent to linear interpolation between B and R_1 which for large values of p, where B and R_1 are many distance intervals apart, may be very inaccurate. In fact, the foregoing seems to suggest that implicit schemes in general are poor, particularly for large values of p.

Difference System III (*Central Difference Formula*). This is given by

$$(6) \qquad u_{r,s+1} = u_{r,s-1} - pu_{r+1,s} + pu_{r-1,s} ,$$

and has been used with success by Malkus and Witt [5] to solve some problems in meteorology involving transport of temperature and vorticity in two dimensions. It is obtained by replacing $\frac{\partial u}{\partial t}$ by $\frac{1}{2\Delta t}(u_{r,s+1} - u_{r,s-1})$ and $\frac{\partial u}{\partial x}$ by $\frac{1}{2\Delta x}(u_{r+1,s} - u_{r-1,s})$. Alternatively, from Figure 3, if GG_1 and PP_1 are the characteristics through G and P respectively, so that $u_G = u_{G_1}$ and $u_P = u_{P_1}$, (6) may be written as

$$u_{P_1} = u_{G_1} - pu_A + pu_C .$$

This result is equivalent to using a parabolic interpolation formula incorporating values of u at A, G_1, and C, thus (6) is expected to be an accurate formula, particularly for small values of $| p |$. In fact, (6) is stable for $-1 \leqq p \leqq 1$, and can only be used if G_1 and P_1 lie between A and C.

It is interesting to compare the foregoing predictions of accuracy with numerical calculations carried out using difference systems I, II, and III in turn to solve (2). Two initial profiles of u are considered, the "roof top" and the "sine" and these are illustrated in Figure 4. All calculations are carried out until a time $\frac{6\Delta x}{V}$ is reached.

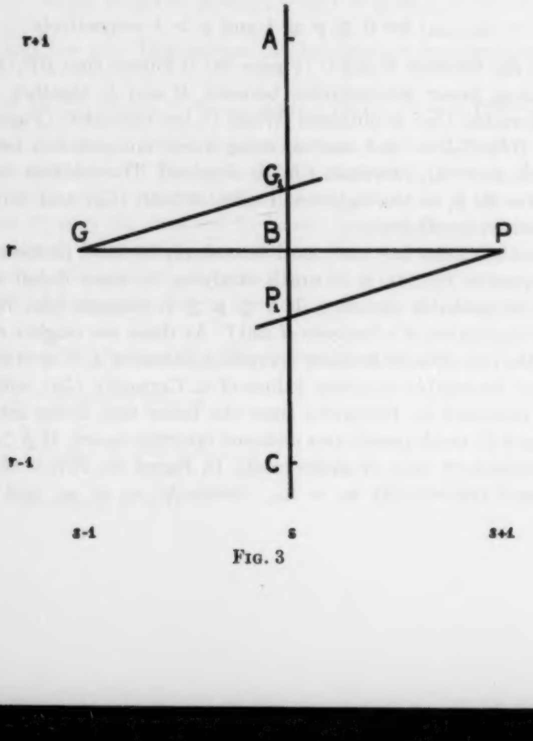


FIG. 3

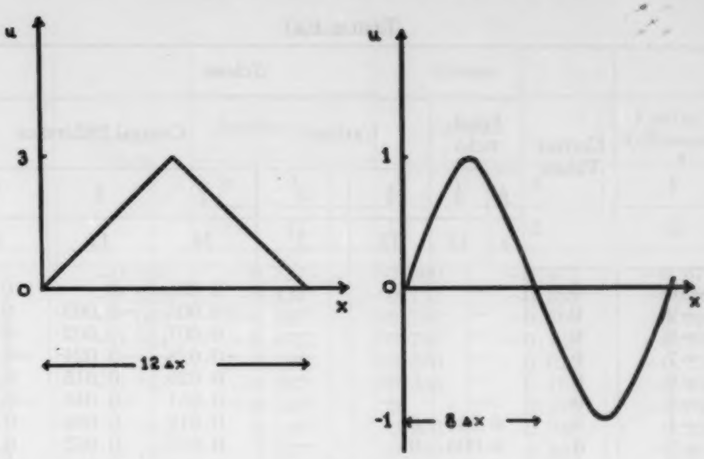


FIG. 4

Theoretical values are used at the second time step in order to start the calculation using System III. The results, accurate to 0.001, are shown in Tables 1(a) and 1(b) for the "roof top" and "sine" profiles respectively. The last row of these tables gives the sum of the moduli of the errors $\sum |e|$. The outstanding features of these results are the poor accuracy of Carlson's scheme for $|p| > 1$, and the comparatively high accuracy of the central difference formula.

3. Two-Level Interpolation Schemes. As a consequence of the last section, explicit difference schemes which are high accuracy interpolation formulas seem most likely to succeed in obtaining accurate solutions of the transport equation. With this in mind, several new two-level formulas are now proposed and used to solve (2). These formulas give $u_{r,s+1}$ in terms of u at nodes on the time step s.

I. Linear Interpolation Formulas.

(7a) $$u_{r,s+1} = (1 - p)u_{r,s} + pu_{r-1,s} \qquad (0 \leqq p \leqq 1)$$

(7b) $$u_{r,s+1} = (2 - p)u_{r-1,s} + (p - 1)u_{r-2,s} \qquad (1 \leqq p \leqq 2)$$

(7c) $$u_{r,s+1} = (n + 1 - p)u_{r-n,s} + (p - n)u_{r-n-1,s} \quad (n \leqq p \leqq n + 1).$$

These formulas are obtained in the following manner using Figure 5. If PP_1 is the characteristic through P so that $u_P = u_{P_1}$ and P_1 lies between B and C, then $BP_1:P_1C = p:1 - p$, and by using linear interpolation of u between B and C, formula (7a) is obtained. If the characteristic through P cuts the line s in P_2 between C and D, then $CP_2:P_2D = p - 1:2 - p$, and linear interpolation of u between C and D gives formula (7b). A similar method may be used for values of p greater than 2.

The general result for any value of p lying between integers n and $n + 1$, where n may be positive or negative, is given by (7c). The formulas are stable since the coefficients on the right-hand sides are positive and add to unity in corresponding pairs.

TABLE 1(a)

r	Correct Values		Scheme					
			Fried-richs	Carlson		Central Difference		
		p	$\frac{1}{2}$	$\frac{1}{2}$	3	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
		s	12	12	2	24	12	8
−10	0		—	—	—	0.001	0	0
−9	0		—	—	—	−0.005	−0.003	0
−8	0		—	—	—	0.007	0.002	0
−7	0		—	—	—	−0.028	−0.024	−0.016
−6	0		—	—	—	0.029	0.015	0.005
−5	0		—	—	—	−0.051	−0.045	−0.033
−4	0		0.001	0	—	0.012	0.006	0.001
−3	0		0.003	0	—	0.053	0.052	0.047
−2	0		0.010	0	—	−0.059	−0.032	−0.010
−1	0		0.017	0	—	0.027	0.006	−0.024
0	0		0.044	0	0.032	0.008	0.014	0.008
1	0		0.071	0.001	0.110	0.166	0.166	0.153
2	0		0.147	0.011	0.241	−0.105	−0.067	−0.024
3	0		0.223	0.047	0.424	−0.270	−0.267	−0.239
4	0		0.384	0.144	0.657	0.021	−0.022	−0.022
5	0		0.545	0.338	0.934	0.200	0.213	0.215
6	0.5		0.796	0.644	1.188	0.465	0.490	0.515
7	1.0		1.046	1.044	1.382	0.842	0.816	0.793
8	1.5		1.311	1.488	1.498	1.676	1.620	1.546
9	2.0		1.575	1.906	1.532	2.403	2.388	2.334
10	2.5		1.715	2.210	1.487	2.605	2.592	2.577
11	3.0		1.856	2.323	1.368	2.631	2.652	2.708
12	2.5		1.773	2.210	1.216	2.426	2.413	2.419
13	2.0		1.691	1.906	1.055	2.023	2.050	2.086
14	1.5		1.428	1.488	0.898	1.408	1.436	1.477
15	1.0		1.166	1.044	0.753	0.817	0.828	0.856
16	0.5		0.873	0.644	0.625	0.424	0.440	0.456
17	0		0.580	0.338	0.514	0.169	0.152	0.116
18	0		0.385	0.144	0.419	0.070	0.066	0.050
19	0		0.190	0.047	0.339	0.019	0.012	0
20	0		0.110	0.011	0.274	0.006	0.004	0
21	0		0.031	0.001	0.219	0	0	0
22	0		0.015	0	0.175	—	—	—
23	0		0	0	0.139	—	—	—
24	0		0	0	0.110	—	—	—
25	0		0	0	0.087	—	—	—
26	0		0	0	0.069	—	—	—
27	0		0	0	0.054	—	—	—
28	0		—	—	0.042	—	—	—
29	0		—	—	0.033	—	—	—
30	0		—	—	0.026	—	—	—
31	0		—	—	0.020	—	—	—
$\sum \lvert e \rvert$			7.298	2.907	12.308	3.000	2.723	2.312

TABLE 1(b)

r	Correct Values	Scheme			
		Friedrichs	Carlson		Central Difference
	p	$\frac{1}{2}$	$\frac{1}{2}$	3	$\frac{1}{2}$
	s	12	12	2	12
0	−0.707	−0.317	−0.560	−0.477	−0.671
1	−0.924	−0.439	−0.731	−0.388	−0.897
2	−1.000	−0.495	−0.792	−0.269	−0.996
3	−0.924	−0.475	−0.731	−0.128	−0.938
4	−0.707	−0.383	−0.560	0.022	−0.738
5	−0.383	−0.232	−0.303	0.162	−0.430
6	0	−0.046	0	0.276	−0.047
7	0.383	0.146	0.303	0.346	0.350
8	0.707	0.317	0.560	0.364	0.671
9	0.924	0.439	0.732	0.327	0.897
10	1.000	0.495	0.792	0.242	0.996
11	0.924	0.475	0.732	0.121	0.938
12	0.707	0.383	0.560	−0.016	0.738
13	0.383	0.232	0.303	−0.150	0.430
14	0	0.046	0	−0.261	0.047
15	−0.383	−0.146	−0.303	−0.330	−0.350
$\sum \lvert e \rvert$		5.174	2.094	7.950	0.478

II. Parabolic Interpolation Formulas.

$$(8)\quad u_{r,s+1} = -\tfrac{1}{2}(p-n)(n+1-p)u_{r+1-n,s} + (p-n+1)(n+1-p)u_{r-n,s} + \tfrac{1}{2}(p-n)(p-n+1)u_{r-1-n,s} \qquad (n \leqq p \leqq n+1).$$

Referring again to Figure 5, if PP_1 is the characteristic through P, where P_1 lies between B and C, and if a parabolic interpolation formula incorporating the values of u at A, B and C is used to give u at points between B and C then $u_{r,s+1}$ is given by (8) with $n = 0$. If the characteristic through P cuts the line s at P_2 where P_2 lies between C and D, and a parabolic interpolation formula incorporating the values of u at B, C, and D is used to give u at points between C and D, then $u_{r,s+1}$ is given by (8) with $n = 1$, and so on for higher values of n. Finally, the stability of (8) is easily demonstrated by using methods described in Richtmyer [1], since the equations are linear and have constant coefficients. Other stable parabolic interpolation schemes based on (8) are possible but they are unlikely to be more accurate than (8) with the original range of p stated.

III. Cubic Interpolation Formulas.

$$\begin{aligned}(9)\quad u_{r,s+1} = &-\tfrac{1}{6}(p-n)(n+1-p)(n+2-p)u_{r+1-n,s}\\ &+ \tfrac{1}{2}(n+2-p)(n+1-p)(p+1-n)u_{r-n,s}\\ &+ \tfrac{1}{2}(p-n)(n+2-p)(p+1-n)u_{r-1-n,s}\\ &- \tfrac{1}{6}(p-n)(n+1-p)(p+1-n)u_{r-2-n,s} \quad (n \leqq p \leqq n+1).\end{aligned}$$

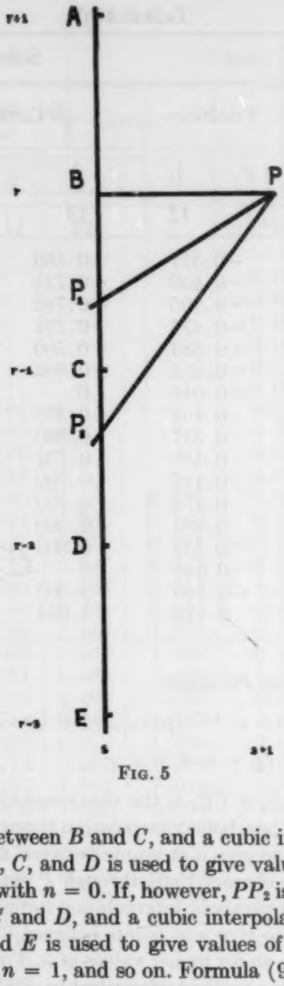


Fig. 5

From Figure 5, if P_1 lies between B and C, and a cubic interpolation formula based on the values of u at A, B, C, and D is used to give values of u between B and C, then $u_{r,s+1}$ is given by (9) with $n = 0$. If, however, PP_2 is the characteristic through P where P_2 lies between C and D, and a cubic interpolation formula based on the values of u at B, C, D and E is used to give values of u between C and D, then $u_{r,s+1}$ is given by (9) with $n = 1$, and so on. Formula (9) is stable not only for the range of p stated but for the extended range $n - 1 \leqq p \leqq n + 2$, and so other cubic interpolation schemes based on (9) are possible. One other possible scheme is (9) together with $n - 1 \leqq p \leqq n + 2$ for $n = \cdots -6, -3, 0, 3, 6, \cdots$. It is unlikely, however, that any of the other schemes will be as accurate as (9) with the original range of p stated and n any integer.

4. Three-Level Formulas. So far the only three-level scheme discussed is the central difference formula. This gives $u_{r,s+1}$ in terms of u at nodes on the time steps $s - 1$ and s. Other three-level formulas suitable for limited ranges of values of p are now proposed.

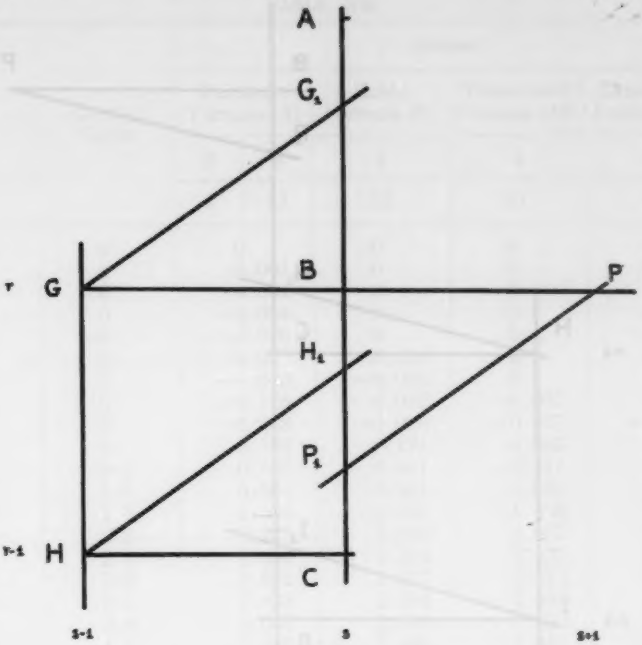


FIG. 6

Referring to Figure 6, PP_1, GG_1 and HH_1 are the characteristics through P, G, and H respectively, thus $u_P = u_{P_1}$, $u_G = u_{G_1}$ and $u_H = u_{H_1}$. If P_1 lies between B and C, and a cubic interpolation formula incorporating the values of u at G_1, B, H_1, and C is used to give the values of u at points between B and C, then the value of u at P is given by

$$u_{r,s+1} = -\frac{(1-2p)(1-p)}{1+p}\, u_{r,s-1} + 2(1-2p)u_{r,s} + 2pu_{r-1,s-1} - \frac{2p(1-2p)}{1+p}\, u_{r-1,s}. \tag{10}$$

In Figure 7, PP_1, HH_1, and II_1 are the characteristics through P, H, and I respectively, thus $u_P = u_{P_1}$, $u_H = u_{H_1}$, and $u_I = u_{I_1}$. If P_1 lies between B and C, and a cubic interpolation formula incorporating the values of u at B, H_1, C and I_1 is used to give the values of u at points between B and C, then the value of u at P is given by

$$u_{r,s+1} = -\frac{2(2p-1)(1-p)}{2-p}\, u_{r,s} + 2(1-p)u_{r-1,s-1} + 2(2p-1)u_{r-1,s} - \frac{p(2p-1)}{2-p}\, u_{r-2,s-1}. \tag{11}$$

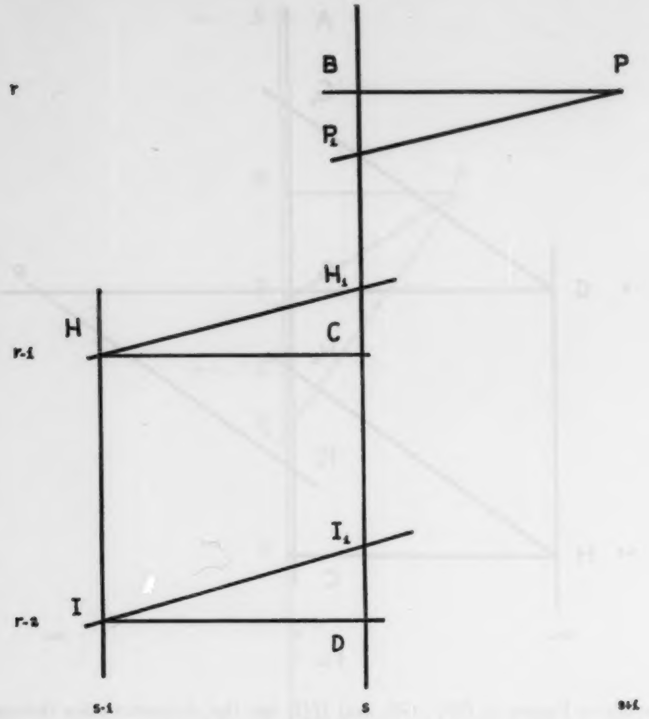


FIG. 7

The stability of (10) and (11) for the range $0 \leqq p \leqq 1$ can be demonstrated in the usual manner.

Numerical calculations are now carried out using selected two- and three-level interpolation schemes to solve (2). The results are shown in Tables 2(a) and 2 (b). The errors are shown in the last two rows where $\sum | e |$ is the sum of the moduli of the errors after a time $6 \frac{\Delta x}{V}$ and $\sum | e_1 |$ refers to the errors at a later stage in the computation when the profile has been transported over a further time $6 \frac{\Delta x}{V}$. In the case of the three-level formula (11), after a time $36 \frac{\Delta x}{V}$ the sums of the moduli of the errors are still only 0.660 and 0.026 for the "roof top" and "sine" curves respectively. The results shown in Tables 2(a) and 2(b) are for values of p lying between 0 and 1, but in the case of the two-level schemes they may be interpreted for values of p outside this range. For example, the figures for $p = \frac{1}{2}$ refer also to $p = n + \frac{1}{2}$ if the profile is moved on a further $12n$ intervals of x.

5. Interpolation Formulas and Finite Difference Schemes. In view of the form of the transport equation, a close link might be expected between interpolation

TABLE 2(a)

		Scheme			
r	Correct Values	Parabolic Formula (8)	Cubic Formula (9)	Three-Level Formula (10)	Three-Level Formula (11)
		p $\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
		s 12	12	24	8
−4	0	0	0	0	0
−3	0	0.001	0	0	0
−2	0	−0.004	0	0	0
−1	0	−0.004	0	0	0
0	0	0.020	0	0	0
1	0	0.024	0.003	0	0
2	0	−0.045	−0.005	0	0
3	0	−0.126	−0.032	−0.003	0.001
4	0	−0.068	−0.026	−0.027	−0.011
5	0	0.165	0.121	0.063	0.038
6	0.5	0.502	0.471	0.471	0.479
7	1.0	0.961	0.961	1.003	1.000
8	1.5	1.591	1.505	1.498	1.499
9	2.0	2.248	2.068	2.007	1.997
10	2.5	2.651	2.554	2.555	2.522
11	3.0	2.682	2.757	2.872	2.923
12	2.5	2.433	2.554	2.557	2.541
13	2.0	2.006	2.068	1.992	1.999
14	1.5	1.452	1.505	1.501	1.500
15	1.0	0.876	0.961	0.996	1.001
16	0.5	0.421	0.471	0.472	0.488
17	0	0.156	0.121	0.063	0.038
18	0	0.043	−0.026	−0.028	−0.020
19	0	0.008	−0.032	0.003	0
20	0	0.001	−0.005	0	0
21	0	0	0.003	0	0
22	0	0	0	0	0
23	0	0	0	0	0
24	0	0	0	0	0
25	0	0	0	0	0
26	0	0	0	0	0
$\sum \lvert e \rvert$		1.838	1.007	0.509	0.287
$\sum \lvert e_1 \rvert$		2.644	1.169	0.660	0.432

formulas and difference schemes used to solve (2). This is best illustrated by means of an example. Consider the problem of evolving a finite difference replacement of (2) which makes use of the points *P*, *B*, *C*, and *D* in Figure 5. Taylor expansions about the point *B* give

$$u_{r,s+1} = u_{r,s} + \Delta t \left(\frac{\partial u}{\partial t}\right)_{r,s} \tag{12}$$

$$u_{r-1,s} = u_{r,s} - \Delta x \left(\frac{\partial u}{\partial x}\right)_{r,s} + \tfrac{1}{2} (\Delta x)^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_{r,s} \tag{13}$$

TABLE 2(b)

r	Correct Values		Scheme			
			Parabolic Formula (8)	Cubic Formula (9)	Three-Level Formula (10)	Three-Level Formula (11)
		p	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
		s	12	12	24	8
0	−0.707		−0.733	−0.702	−0.706	−0.706
1	−0.924		−0.933	−0.917	−0.922	−0.923
2	−1.000		−0.992	−0.993	−0.999	−0.999
3	−0.924		−0.900	−0.917	−0.923	−0.923
4	−0.707		−0.670	−0.702	−0.706	−0.706
5	−0.383		−0.338	−0.380	−0.382	−0.382
6	0		0.044	0	0	0
7	0.383		0.420	0.380	0.382	0.382
8	0.707		0.733	0.702	0.706	0.706
9	0.924		0.933	0.917	0.922	0.923
10	1.000		0.992	0.993	0.999	0.999
11	0.924		0.900	0.917	0.923	0.923
12	0.707		0.670	0.702	0.706	0.706
13	0.383		0.338	0.380	0.382	0.382
14	0		−0.044	0	0	0
15	−0.383		−0.420	−0.380	−0.382	−0.382
$\sum \mid e \mid$			0.460	0.074	0.016	0.014
$\sum \mid e_1 \mid$			0.914	0.144	0.026	0.014

and

$$(14) \qquad u_{r-2,s} = u_{r,s} - 2\Delta x \left(\frac{\partial u}{\partial x}\right)_{r,s} + 2(\Delta x)^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_{r,s}.$$

The value of $\left(\frac{\partial u}{\partial t}\right)_{r,s}$ is obtained from (12) and $\left(\frac{\partial u}{\partial x}\right)_{r,s}$ by eliminating $\left(\frac{\partial^2 u}{\partial x^2}\right)_{r,s}$ from (13) and (14). These values are then substituted into (2) to give

$$(15) \qquad u_{r,s+1} = \left(1 - \frac{3p}{2}\right) u_{r,s} + 2pu_{r-1,s} - \frac{p}{2} u_{r-2,s}.$$

The truncation error in (15) is dominated by the term $\frac{1}{2}(\Delta t)^2 \left(\frac{\partial^2 u}{\partial t^2}\right)_{r,s}$ neglected in (12) and since by differentiating (2) the result

$$(16) \qquad \frac{\partial^2 u}{\partial t^2} = V^2 \frac{\partial^2 u}{\partial x^2}$$

is obtained, it follows that the principal part of the truncation error is $\frac{1}{2}p^2(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}$. This is the standard finite difference approach which can, however, be improved in the following manner. Replace equation (12) by

$$(17) \qquad u_{r,s+1} = u_{r,s} + \Delta t \left(\frac{\partial u}{\partial t}\right)_{r,s} + \frac{1}{2} (\Delta t)^2 \left(\frac{\partial^2 u}{\partial t^2}\right)_{r,s},$$

which, on using (2) and (16) becomes

$$u_{r,s+1} = u_{r,s} - p\Delta x \left(\frac{\partial u}{\partial x}\right)_{r,s} + \tfrac{1}{2}\, p^2(\Delta x)^2 \left(\frac{\partial^2 u}{\partial x^2}\right)_{r,s}. \tag{18}$$

If $\left(\frac{\partial u}{\partial x}\right)_{r,s}$ and $\left(\frac{\partial^2 u}{\partial x^2}\right)_{r,s}$ are now eliminated from (13), (14), and (18), the parabolic interpolation formula (8) with $n = 1$ is obtained with truncation error

$$\tfrac{1}{6}p(1 - p)(2 - p)(\Delta x)^3 \frac{\partial^3 u}{\partial x^3}.$$

This is a distinct improvement over the previous finite difference formula (15), and in particular if p is close to unity, the interpolation formula is expected to be specially accurate when used to solve (2). If $p = 1$, of course, the theoretical solution of the interpolation formula (8) with $n = 1$ is the same as the theoretical solution of (2). However, as it is intended to use the results of the present investigation to solve the general transport equation (1), the exact correspondence of the theoretical solutions of (2) when $p = 1$ can really be ignored. This example illustrates the fact that the best finite difference formula for a given set of points used to solve (2) is an interpolation formula. This is because each derivative with respect to a

TABLE 3

Scheme	*Formula Number*	*Truncation Error*
Friedrichs	(4)	$\frac{1}{2}(1 - p^2)(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}$
Carlson $0 \leqq p \leqq 1$	(5a)	$\frac{1}{2}p(1 - p)(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}$
Carlson $p > 1$	(5b)	$\frac{1}{2}p(p + 1)(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}$
Central Difference	(6)	$-\frac{1}{6}p(1 - p^2)(\Delta x)^3 \frac{\partial^3 u}{\partial x^3}$
Linear Interpolation	(7c)	$-\frac{1}{2}(n - p)(n - p + 1)(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}$
Parabolic Interpolation	(8)	$-\frac{1}{6}(n - p - 1)(n - p)(n - p + 1)(\Delta x)^3 \frac{\partial^3 u}{\partial x^3}$
Cubic Interpolation	(9)	$\frac{1}{24}(n - p - 1)\,(n - p)\,(n - p + 1)$ $\cdot(n - p + 2)(\Delta x)^4 \frac{\partial^4 u}{\partial x^4}$
Three-Level I	(10)	$-\frac{1}{2}p^2(1 - p)(1 - 2p)(\Delta x^4) \frac{\partial^4 u}{\partial x^4}$
Three-Level II	(11)	$\frac{1}{2}p(1 - p)^2(1 - 2p)(\Delta x)^4 \frac{\partial^4 u}{\partial x^4}$

co-ordinate is a constant multiple of the corresponding derivative with respect to the other co-ordinate, thus the Taylor expansions can all be expressed in terms of a single variable. Elimination of the maximum possible number of derivatives with respect to this variable leads to an interpolation formula.

6. Truncation Errors. For purposes of comparison, the truncation errors associated with the finite difference schemes considered for solving (2) are given in Table 3. The errors quoted are Δx times the errors as defined by Richtmyer [1, p. 19].

7. Conclusions. The calculations carried out in the present paper, using existing stable finite difference schemes in turn to solve the simplified transport equation (2), vary considerably in accuracy. The central difference formula (6) is most accurate with Carlson's scheme for $| p | \leqq 1$ next in order of merit. Carlson's implicit scheme for $| p | > 1$ is very poor, particularly for large values of $| p |$. This is illustrated in Figure 8 where the part of the truncation error depending on p is shown as a function (E) of p. It can be seen that the maximum value of the truncation error when $0 \leqq p \leqq 1$ is one-eighth of the minimum value when $p > 1$. In fact, the authors believe that implicit schemes can be abandoned as a means of obtaining accurate solutions of the transport equation.

New explicit schemes, derived as interpolation formulas, are next used to solve (2) and a considerable improvement in accuracy is obtained, particularly for schemes such as (9), (10), and (11), which are cubic interpolation formulas with a very small truncation error. The error in any numerical solution of (2) takes the form of a smoothing out of the initial profile together with, in most cases, a superposed stable oscillation.

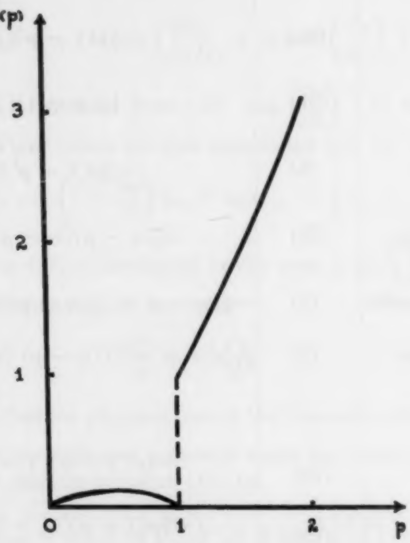


Fig. 8

It cannot be emphasized too strongly, however, that schemes which successfully solve (2) do not necessarily give comparable accuracy when used to solve (1), where V is a function of x, t, and u. On the other hand, difference schemes which fail to give accurate solutions of (2), can hardly be expected to be more successful when used to solve (1). The main difficulty in solving (1) numerically arises from the fact that the characteristics are curved and the distance BP_1 (Figures 1, 2, 3, 5, 6, 7) is no longer given simply by $V\Delta t$ or $p\Delta x$. It must be found by integrating the equation

$$\frac{dx}{dt} - V(x, t, u) = 0. \tag{19}$$

If, in the case of curved characteristics, BP_1 is now expressed as $p'\Delta x$, any one of the interpolation formulas proposed in the present paper may be applied directly with p' substituted for p. The value of p' is, of course, in general different at each node.

In deciding the values of Δx and Δt for a given calculation, Δx is first chosen to represent adequately the initial profile. The time step Δt is then chosen so that BP_1 is large enough for the calculation to proceed without too many interpolations but not so large that the positions of P_1, obtained from (19), are too much in error. We hope to examine in detail at a later date the general problem of integrating (1).

8. Acknowledgements. The calculations were performed on the Deuce computer of Glasgow University, and the authors are indebted to Dr. D. C. Gillies, Director of the Computing Laboratory, for permission to use the machine and to Miss A. Low for help in running the programs.

Mathematics Department
Queen's College
Dundee, Scotland and

Mathematics Department
St. Salvat<r's College
St. Andrews, Scotland

1. R. D. Richtmyer, *Difference Methods for Initial-Value Problems*, Interscience Publishers Inc., New York, 1957.

2. K. O. Friedrichs, "Symmetric hyperbolic linear differential equations," *Comm. Pure Appl. Math.*, v. 7, 1954, p. 345–392.

3. B. G. Carlson, Unpublished Los Alamos Report, 1953.

4. H. B. Keller & B. Wendroff, "On the formulation and analysis of numerical methods for time-dependent transport equations," *Comm. Pure Appl. Math.*, v. 10, 1957, p. 567–582.

5. J. S. Malkus & G. Witt, "The atmosphere and the sea in motion," *Rossby Memorial Volume*, Rockefeller Institute Press, New York, 1959, p. 425.

Quadrature Formulas for Infinite Integrals

By W. M. Harper

1. Introduction. Since the advent of high-speed computers, "mechanical" quadratures of the type

$$(1) \qquad \int_a^b w(x)f(x)\,dx \sim \sum_{j=1}^n H_j\,f(a_j)$$

have become increasingly important. The only quadrature generally available for the case $b = -a = \infty$ is the Hermite-Gauss formula although the Laguerre-Gauss formula can also be used if $f(x)$ is an even function of x. The latter would, however, require computation of twice the number of ordinates for a corresponding degree of precision and would therefore rarely be preferred. In either case the integrand is supposed to behave like the product of an exponential function and a polynomial. For purely algebraic integrands it would appear to be more appropriate to use a quadrature based on an algebraic weight function even though the degree of the polynomial approximation to $f(x)$ is limited.

In this paper, formulas of type (1) are derived with weight function $w(x) = (1 + x^2)^{-k-1}$ for the range $b = -a = \infty$. In a modified form they are shown to be superior to the Hermite-Gauss and Laguerre-Gauss quadratures for a particular class of statistical integrals.

2. Derivation of Quadratures. In the quadrature formula

$$(2) \qquad \int_{-\infty}^{\infty} (1 + x^2)^{-k-1} f(x)\,dx = \sum_{j=1}^n H_j\,f(a_j) + E_{n,k},$$

the abscissas a_j will be the zeros of the nth degree polynomial $\phi_{n,k}(x)$ which satisfies the orthogonality condition

$$(3) \qquad \int_{-\infty}^{\infty} (1 + x^2)^{-k-1} \phi_{m,k}(x)\phi_{n,k}(x)\,dx = 0, \qquad (m \neq n, m + n < 2k + 1).$$

By standard methods given for example in [2], [4], it is easily shown from (3) that the orthogonal system of polynomials is given by the Rodrigues formula

$$(4) \quad \phi_{n,k}(x) = (-1)^n \frac{\Gamma(2k - 2n + 2)}{\Gamma(2k - n + 2)} (1 + x^2)^{k+1} \frac{d^n}{dx^n} (1 + x^2)^{n-k-1}, \qquad (n < k + 1)$$

where the standardizing constant is chosen to make the coefficient of x^n unity. By direct manipulations with (4) and repeated use of Leibnitz' formula, the recurrence relations (5)–(10) are easily established. They are

$$(5) \qquad \phi_{n+1,k}(x) = x\phi_{n,k}(x) - \frac{n(2k - n + 2)}{(2k - 2n + 1)(2k - 2n + 3)} \phi_{n-1,k}(x),$$

$$(6) \quad (1 + x^2)\phi'_{n,k}(x) = (2k - n + 1)x\phi_{n,k}(x) - (2k - 2n + 1)\phi_{n+1,k}(x),$$

Received August 28, 1961.

$$(1 + x^2)\phi'_{n,k}(x) = nx\phi_{n,k}(x) + \frac{n(2k - n + 2)}{2k - 2n + 3}\phi_{n-1,k}(x), \tag{7}$$

$$\begin{aligned}\phi_{n,k+1}(x) = {} & \frac{2k - 2n + 3}{(2k - n + 2)(2k - n +3)} \\ & \cdot[\{(4k - 2n + 3) + (2k - 2n + 1)x^2\}\,\phi_{n,k}(x) \\ & \qquad - (2k - 2n + 1)(1 + x^2)\phi_{n,k-1}(x)],\end{aligned} \tag{8}$$

$$\begin{aligned}x(1 + x^2)\phi'_{n,k}(x) = {} & [nx^2 - (2k - n + 2)]\phi_{n,k}(x) \\ & + \frac{(2k - n + 2)(2k - n + 3)}{2k - 2n + 3}\phi_{n,k+1}(x),\end{aligned} \tag{9}$$

$$x\phi'_{n,k}(x) = (2k - n + 1)\phi_{n,k}(x) - (2k - 2n + 1)\phi_{n,k-1}(x). \tag{10}$$

The polynomial system can now be extended to include values of n excluded in (4). For $n > k + \frac{3}{2}$ however, complex zeros make their appearance so that no useful quadratures are available for this range of n.

It is similarly easily shown that $\phi_{n,k}(x)$ is a solution of the differential equation

$$(1 + x^2)y'' - 2kxy' + n(2k - n + 1)y = 0 \tag{11}$$

whence the relation

$$\phi_{n,k}(x) = \left(\frac{i}{2}\right)^n n!\,\frac{\Gamma(k - n + \frac{3}{2})}{\Gamma(k + \frac{3}{2})}\,C_n^{-k-1/2}(ix) \tag{12}$$

can be established where in the notation of [1], $C_n^{\lambda}(z)$ (designated by $P_n^{(\lambda)}(z)$ in [7]) is the Gegenbauer or ultraspherical polynomial of degree n and parameter λ. Relations with Legendre functions can also be established, namely:

$$\begin{aligned}\phi_{n,k}(x) = {} & (-1)^{n+1}\pi^{1/2}\lim_{s\to k} \\ & \cdot\left[2^{s-n}i^{s+n+1}\frac{\Gamma(s - n + \frac{3}{2})}{\Gamma(2s - n + 2)}\operatorname{cosec} s\pi(1 + x^2)^{s/2+1/2}P_{s-n}^{s+1}(ix)\right]\end{aligned} \tag{13}$$

where $P_\nu^\mu(z)$, in the notation of [1], is the associated Legendre function of the first kind with parameters μ and ν, and

$$\phi_{n,k}(x) = 2^{k-n+1/2}\Gamma(k - n + \tfrac{3}{2})(1 + x^2)^{k/2+1/4}\mathrm{P}_{k+1/2}^{-k+n-1/2}[x(1 + x^2)^{-1/2}] \tag{14}$$

where $\mathrm{P}_\nu^\mu(z)$ is the associated Legendre function of the first kind with definition suitable for the cut in the real axis from $z = -1$ to $z = 1$. The limit in (13) caters to integer values of k (see [3]).

The weight coefficients and error term in (2) can be determined by standard methods with the results

$$H_j = 2^{2k-2n+2}n!\,\frac{[\Gamma(k - n + \frac{3}{2})]^2}{\Gamma(2k - n + 2)}(1 + a_j^2)^{-1}[\phi'_{n,k}(a_j)]^{-2}, \tag{15}$$

$$\left\{\begin{aligned}E_{n,k} &= \frac{f^{(2n)}(\xi)}{(2n)!}\int_{-\infty}^{\infty}(1 + x^2)^{-k-1}[\phi_{n,k}(x)]^2\,dx \\ &= \frac{2^{2k-2n+2}n![\Gamma(k - n + \frac{3}{2})]^2}{(2k - 2n + 1)(2n)!\,\Gamma(2k - n + 2)}f^{(2n)}(\xi), \qquad \left(n < k + \frac{1}{2}\right).\end{aligned}\right. \tag{16}$$

The restriction on n is necessary to ensure convergence of the error estimate but does not ensure that a close upper bound to the actual error can be obtained (see, for example, [2]).

For practical purposes a more convenient form of the quadrature is

$$\int_{-\infty}^{\infty} f(x)\,dx \sim \sum_{j=1}^{n} K_j f(a_j); \tag{17}$$

here the weight coefficients are given by

$$K_j = H_j(1 + a_j^2)^{k+1}. \tag{18}$$

The values of a_j and K_j for four- and six-point formulas for some integral values of k are given in Table 1.

The right-hand side of (17) is a function of k as well as of n; for a given value of n, therefore, there will be a value or values of k depending on $f(x)$ which will give the "best" approximation to the integral on the left. The determination of such values and the corresponding parameters appears to be too formidable a task for practical applications. For the special cases $k = n - 1$, $k = n$, however, solution of (11) with $x = \cot\theta$ enables $\phi_{n,k}(x)$ to be obtained in the forms

$$\phi_{n,n-1}(x) = \operatorname{cosec}^n(\operatorname{arc\,cot} x)\cos(n\operatorname{arc\,cot} x), \tag{19}$$

$$\phi_{n,n}(x) = (n+1)^{-1}\operatorname{cosec}^{n+1}(\operatorname{arc\,cot} x)\sin[(n+1)\operatorname{arc\,cot} x]. \tag{20}$$

The zeros are now simple cotangents and the weight coefficients H_j assume simple trigonometric form; the resulting quadratures can be written as

$$\int_{-\infty}^{\infty}(1+x^2)^{-1}f(x)\,dx \sim \frac{\pi}{n}\sum_{j=1}^{n} f\left[\cot\frac{(2j-1)\pi}{2n}\right], \qquad (k = n-1), \tag{21}$$

$$\int_{-\infty}^{\infty}(1+x^2)^{-1}f(x)\,dx \sim \frac{\pi}{n+1}\sum_{j=1}^{n} f\left(\cot\frac{j\pi}{n+1}\right), \qquad (k = n). \tag{22}$$

These formulas can also be deduced from the Chebyshev-Gauss quadratures

$$\int_{-1}^{1}(1-y^2)^{-1/2}g(y)\,dy \sim \frac{\pi}{n}\sum_{j=1}^{n} g\left[\cos\frac{(2j-1)\pi}{2n}\right], \tag{23}$$

$$\int_{-1}^{1}(1-y^2)^{-1/2}g(y)\,dy \sim \frac{\pi}{n+1}\sum_{j=1}^{n} g\left(\cos\frac{j\pi}{n+1}\right) \tag{24}$$

by the substitutions $y = x(1+x^2)^{-1/2}$, $g(y) = f(x)$.

3. Practical Application. An example of a useful application for the quadratures is the evaluation of integrals arising in the determination of the statistical distribution of the ratio of two quadratic forms in normal variates. If the quadratic forms are independent mean half-square successive differences based on sample sizes of p and q respectively, one of the integrals which require evaluation can be written in the form

$$I(z) = \int_{-\infty}^{\infty}(1+x^2)^{-1}\prod_{r=2}^{p-1}(a_r^2+x^2)^{-1/2}\prod_{s=1}^{q-1}(1+b_s^2z^{-1}+x^2)^{-1/2}\,dx, \qquad (p \text{ even}), \tag{25}$$

Table 1
Abscissas and Weights for Quadrature (17)

A. $n = 4$

k	$\pm a_j$	K_j
3	0.41421 35624	0.92015 11845
	2.41421 35624	5.36303 41227
4	0.32491 96962	0.69465 18830
	1.37638 19205	1.81862 22399
5	0.27618 30252	0.58086 65620
	1.06005 79874	1.17945 11502
6	0.24436 83118	0.50932 47880
	0.89298 76737	0.90816 46087
7	0.22150 78137	0.45903 94023
	0.78587 59159	0.75578 97944
8	0.20405 97869	0.42121 27662
	0.70979 86678	0.65698 70999
9	0.19017 76238	0.39142 46836
	0.65220 46710	0.58705 73261
10	0.17879 14705	0.36717 90805
	0.60665 77372	0.53455 96626

B. $n = 6$

k	$\pm a_j$	K_j
5	0.26794 91924	0.56119 14763
	1.00000 00000	1.04719 75512
	3.73205 08076	7.81638 89333
6	0.22824 34744	0.47217 91694
	0.79747 33889	0.73421 88392
	2.07652 13966	2.38399 35955
7	0.20219 80919	0.41550 76425
	0.68370 47228	0.58969 00381
	1.57850 04858	1.44716 80133
8	0.18342 80037	0.37535 93234
	0.60816 30047	0.50404 67421
	1.31884 38384	1.06492 43997
9	0.16907 35256	0.34499 40643
	0.55326 32106	0.44635 57833
	1.15411 46518	0.85743 60559
10	0.15763 63749	0.32098 68394
	0.51101 94490	0.40432 69556
	1.03809 74230	0.72680 65190

TABLE 2

Comparison of Quadrature Formulas in Evaluating $I(1)$

Quadrature	No. Abscissas	Result	$E \times 10^8$
Series	—	1.2106 5423	—
Algebraic, $k = 5$	6	1.2106 4384	1039
Algebraic, $k = 6$	6	1.2106 5381	42
Algebraic, $k = 7$	6	1.2106 5415	8
Algebraic, $k = 8$	6	1.2081 0423	25 5000
Algebraic, $k = 9$	6	1.2025 0816	81 4607
Algebraic, $k = 10$	6	1.1942 4044	164 1379
Hermite	6	1.1610 8623	495 6800
Hermite	8	1.1879 0738	227 4685
Hermite	10	1.1994 3337	112 2086
Laguerre	6	1.1674 2007	432 3416

where the a_r and b_s are constants. In order to compare methods (25) was evaluated by various quadratures for the case $p = 4$, $q = 3$, $z = 1$ when the test integral becomes

$$I(1) = \int_{-\infty}^{\infty} (1 + x^2)^{-1} \left[\left(\frac{1}{2} \sqrt{2} + x^2 \right) (2\sqrt{2} - 2 + x^2) \right. \tag{26}$$

$$\left. \cdot \left\{ \frac{1}{3} (7 - 2\sqrt{2}) + x^2 \right\} \left\{ \frac{1}{9} (13 - 2\sqrt{2}) + x^2 \right\} \right]^{-1/2} dx.$$

The quadrature (17) was applied for the values $k = 5(1)10$ using six abscissas in each case. The Hermite-Gauss quadrature was used with six, eight and ten abscissas, and the Laguerre-Gauss formula for six abscissas (which requires the same number of evaluations of the integrand as the other formulas for twelve abscissas but which is of degree of precision eleven as against twenty-three for the others). The abscissas and weights for the Hermite formula were taken from the values tabulated in [6] and those for the Laguerre method from [5]. The results together with the correct value of $I(1)$ determined by a series method are tabulated to eight decimal places in Table 2 which also shows the errors of the methods.

The table shows the superiority of the "algebraic" quadratures over the Hermite and Laguerre formulas for this integral; even the use of ten abscissas for the Hermite quadrature leaves an error much greater than the algebraic quadratures with only six abscissas except for the case $k = 10$. The best algebraic quadrature is for $k = 7$ but the advantage over those for $k = 5$ and $k = 6$ is too small to compensate for the simplicity of the latter two cases when used in the equivalent forms shown in (21) and (22) respectively. In addition, the quadrature (22) evaluates $I(1)$ correctly to eight decimal places for $n = 8$ as does (21) for $n = 9$.

4. Acknowledgements. The author is indebted to Mr. J. A. Macdonald for some helpful suggestions and to Miss Primula Thomson for much of the computing.

The paper is published with the permission of the Chief Scientist, Department of Supply, Australian Defence Scientific Service, Melbourne, Victoria, Australia.

Department of Supply
Australian Defence Scientific Service
Defence Standards Laboratories
Maribyrnong, Victoria, Australia.

1. A. Erdélyi, et al., *Higher Transcendental Functions*, McGraw-Hill, New York, 1953, v. 1, p. 120–181, v. 2, p. 174–178.

2. F. B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1956, p. 312–367.

3. E. W. Hobson, *Spherical and Ellipsoidal Harmonics*, Cambridge University Press, 1931, p. 178–292.

4. Z. Kopal, *Numerical Analysis*, Wiley & Sons, New York, 1955, p. 347–439.

5. H. E. Salzer & R. Zucker, "Table of the zeros and weight factors of the first fifteen Laguerre polynomials," *Bull. Amer. Math. Soc.* v. 55, 1949, p. 1004–1012.

6. H. E. Salzer, R. Zucker & R. Capuano, "Table of the zeros and weight factors of the first twenty Hermite polynomials," *J. Res. Nat. Bur. Standards*, v. 48, 1952, p. 111–116.

7. G. Szegö, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, v. 23, 1959, p. 81–86.

A Modification of the Runge-Kutta Fourth-Order Method

By E. K. Blum

1. Introduction. Consider the system of n first-order ordinary differential equations,

$$y_k' = f_k(t, y_1(t), \cdots, y_n(t)), \qquad k = 1, \cdots, n, \tag{1.1}$$

with the initial values,

$$y_k(t_0) = a_k\,. \tag{1.2}$$

Under suitable conditions on the f_i, a unique solution of (1.1) satisfying (1.2) exists for some interval, $t_0 \leqq t \leqq b$. For example, it is sufficient that the f_i be continuous and satisfy a Lipschitz condition in some neighborhood of the initial point, $(t_0, a_1, \cdots, a_n)$. We shall assume that such conditions obtain, so that the initial value problem (1.1), (1.2) has a unique solution.

To simplify the notation, we define $y_0 \equiv t$ and $f_0 \equiv 1$. We now let y be the vector, $(y_0, y_1, \cdots, y_n)$, and f the vector-valued function, $(f_0, f_1, \cdots, f_n)$. The initial value problem can then be written as

$$y' = f(y), \tag{1.3}$$

$$y(t_0) = a. \tag{1.4}$$

The Runge-Kutta fourth-order method for the numerical solution of (1.3), (1.4) yields approximate values, y_j, of y on a finite set of points, $t_j = t_0 + jh$, $j = 1, 2, \cdots, m$. It is usually summarized in formulas (1.5)–(1.9) below, which specify the calculations to be carried out for each integration step; i.e. for each value of j.

$$k_1 = hf(y_j), \tag{1.5}$$

$$k_2 = hf(y_j + k_1/2), \tag{1.6}$$

$$k_3 = hf(y_j + k_2/2), \tag{1.7}$$

$$k_4 = hf(y_j + k_3), \tag{1.8}$$

$$y_{j+1} = y_j + (k_1 + 2k_2 + 2k_3 + k_4)/6. \tag{1.9}$$

A variant of this method was derived by S. Gill [1]. The two advantages of Gill's variant are (1) in automatic computers, it requires $3n + B$ storage registers whereas the Runge-Kutta formulas as given above, require $4n + B$, where B is some constant; (2) the computation can be arranged so that rounding errors are reduced appreciably. In the present paper, we shall show how, by means of a fairly simple modification of (1.5)–(1.9), both of these advantages can be made to accrue to the classical Runge-Kutta method. All the constants in this modification are rational, whereas Gill's variant contains some irrational constants. The modifica-

Received September 18, 1961.

tion is achieved by extracting from Gill's method its main virtue, the rather ingenious device for reducing the rounding error, and applying it to a rearrangement of (1.5)–(1.9).

2. The Exact Modification. In an automatic digital computer, real numbers are replaced by what von Neumann and Goldstine [2] call "digital numbers," that is, by real numbers rounded to a prescribed number of digits. Further, exact arithmetic operations are replaced by "pseudo-operations" since results must be rounded. The main advantage of the modified Runge-Kutta formulas to be presented in Section 3 is that they reduce considerably the rounding error arising from the unavoidable use of digital numbers and pseudo-operations. The saving of n storage registers is a secondary consideration in large computers. The same is true of the Gill variant.

In this section we shall present a preliminary version of the proposed method. We shall refer to it as the "exact modification" since all operations will be assumed to be exact operations on real numbers. The form of the exact modification will demonstrate clearly how the saving of n storage registers is effected.

Using vector notation, as in (1.3)–(1.9), we can write the exact modification in a recursive form as follows:

$$\begin{cases} z_0 = y_j , \\ q_0 = y_j , \\ P_0 = hf(z_0), \end{cases} \tag{2.1}$$

$$\begin{cases} z_1 = z_0 + P_0/2, \\ q_1 = P_0 , \\ P_1 = hf(z_1), \end{cases} \tag{2.2}$$

$$\begin{cases} z_2 = z_1 + P_1/2 - q_1/2, \\ q_2 = q_1/6, \\ P_2 = hf(z_2) - P_1/2, \end{cases} \tag{2.3}$$

$$\begin{cases} z_3 = z_2 + P_2 , \\ q_3 = q_2 - P_2 , \\ P_3 = hf(z_3) + 2P_2 , \end{cases} \tag{2.4}$$

$$y_{j+1} \equiv z_4 = z_3 + q_3 + P_3/6. \tag{2.5}$$

(Strictly speaking, each of the vectors, z_i , q_i , P_i , should have a second subscript, j, to indicate that the sequence (2.1)–(2.5) is repeated for each step of the solution. This subscript has been dropped for reasons of economy, just as the subscript which indicates the components of the vectors has been dropped.)

THEOREM 1. *The exact modification, (2.1)–(2.5), is equivalent to the classical Runge-Kutta method and requires only $3n + B$ storage registers.*

Proof. To show that (2.1)–(2.5) is equivalent to (1.5)–(1.9), we first observe that $P_0 = k_1$. Then $z_1 = y_j + k_1/2$, which implies $P_1 = k_2$. Since $q_1 = k_1$, it follows that $z_2 = (y_j + k_1/2) + k_2/2 - k_1/2 = y_j + k_2/2$. Thus,

$$P_2 = k_3 - k_2/2$$

and $q_2 = k_1/6$. From (2.4), it now follows that $z_3 = (y_j + k_2/2) + (k_3 - k_2/2) =$

$y_j + k_3$, and $q_3 = k_1/6 - (k_3 - k_2/2)$, whence $P_3 = k_4 + 2k_3 - k_2$. Combining these expressions in (2.5), we get

$$y_{j+1} = (y_j + k_3) + \left(\frac{k_1}{6} - k_3 + \frac{k_2}{2}\right) + \frac{1}{6}(k_4 + 2k_3 - k_2),$$

$$y_{j+1} = y_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

The order of computation of the components of the vectors z_{i+1}, q_{i+1}, and P_{i+1}, $i = 0, 1, 2, 3$, should be as follows. First, compute the components of z_{i+1} and q_{i+1} together. Each such component involves only the corresponding component of z_i, q_i, and P_i. Hence, as each component of z_{i+1} and q_{i+1} is computed, it can be placed in the storage occupied by the corresponding component of z_i and q_i, respectively. After all components of z_{i+1} and q_{i+1} have been computed, the components of P_{i+1} can be computed, replacing the corresponding components of P_i in storage. Thus, $3n + B$ storages suffice.

3. The Finite-Precision Modification. In this section we shall consider the rounding errors which arise in actual computation when digital numbers and pseudo-operations are used in (2.1)–(2.5). We shall adopt the notation of [2] for digital numbers, denoting a digital number by a letter with a bar over it, and similarly for vectors; e.g. $\bar{y}_i$ is a vector having digital numbers as components. However, we shall not introduce special symbols for pseudo-operations. Instead, we prefer to write all formulas with exact operations and introduce special terms to denote the rounding error caused by the pseudo-operations. Besides the usual arithmetic operations, we require two "shifting" operations. These are best described informally.

For the remainder of this section, let us assume that a digital number is represented by a sequence of s decimal digits, and that the decimal point is at the extreme left. (The first digit immediately to the right of the decimal point is said to be in position 1.) For m a non-negative integer, we define the operator, R_m, ("shift right m places") as follows. If $\bar{y}$ is a digital number, then $R_m\bar{y}$ is the digital number obtained by shifting the digits of $\bar{y}$ m positions to the right, "rounding off" the digits shifted into positions $s + 1, \cdots, s + m$, and inserting zeros into positions $1, \cdots, m$. The usual method of rounding off is to add (if $\bar{y} \geqq 0$) or subtract (if $\bar{y} < 0$) the digit "5" in position $s + 1$, and then drop the digits beyond position s. The operator, L_m, ("shift left m places") is defined similarly. Thus, $L_m\bar{y}$ is the digital number obtained by shifting the digits of $\bar{y}$ to the left m places, dropping those digits which are then to the left of the decimal point, and inserting zeros into positions $s - m + 1, \cdots, s$. When applied to vectors, R_m and L_m are considered to operate on each of the component digital numbers. If y is a real number, then we define $R_m y = 10^{-m}y$ and $L_m y = 10^m y$.

To briefly motivate the formulas to be given below, let us consider the exact modification, (2.1)–(2.5). If this procedure were carried out with digital numbers and pseudo-operations, an analysis of the rounding error would show that, under suitable conditions on the partial derivatives, $\partial f_k/\partial y_i$, the main source of error is in the computation of the z_i. The error there arises from the fact that q_i and P_i

are usually smaller than z_i in magnitude, since they are of order h. This means that either q_i and P_i must be computed with fewer significant digits than z_i, or else a shift right must be performed before q_i and P_i can be added to z_i. In either case an appreciable rounding error is incurred. The procedure explained below reduces this particular error. We shall refer to it as the "finite-precision modification" of the Runge-Kutta method to emphasize that it is designed for actual computation with digital numbers having a "finite precision" of s places. The finite-precision modification is derived from the exact modification by introducing the special quantities, r_i, as in Gill's formulas. To facilitate the error analysis, we shall write the finite-precision modification first with real numbers and exact operations, (3.1)–(3.5), and then with digital numbers and error terms, $(\overline{3.1})$–$(\overline{3.5})$.

$$\tag{3.1} \left\{\begin{array}{rl} z_0 &= y_j\,, \\ L_mq_0 &= L_mq_{4j}\,, \\ L_mP_0 &= (L_mh)f(z_0), \end{array}\right.$$

$$\tag{3.2} \left\{\begin{array}{l} \left\{\begin{array}{rl} r_1 &= R_m(\tfrac{1}{2}L_mP_0 - L_mq_0), \\ z_1 &= z_0 + r_1\,, \\ L_mq_1 &= 3L_mr_1 - (\tfrac{1}{2}L_mP_0 - L_mq_0), \end{array}\right. \\ \quad L_mP_1 = (L_mh)f(z_1), \end{array}\right.$$

$$\tag{3.3} \left\{\begin{array}{l} \left\{\begin{array}{rl} r_2 &= R_m(\tfrac{1}{2}(L_mP_1 - L_mq_1)), \\ z_2 &= z_1 + r_2\,, \\ L_mq_2 &= -L_mr_2 - \tfrac{1}{3}L_mq_1 + \tfrac{1}{2}L_mP_1\,, \end{array}\right. \\ \quad L_mP_2 = (L_mh)f(z_2) - \tfrac{1}{2}L_mP_1\,, \end{array}\right.$$

$$\tag{3.4} \left\{\begin{array}{l} \left\{\begin{array}{rl} r_3 &= R_m(L_mP_2), \\ z_3 &= z_2 + r_3\,, \\ L_mq_3 &= -L_mr_3 + L_mq_2\,, \end{array}\right. \\ \quad L_mP_3 = (L_mh)f(z_3) + 2L_mP_2\,, \end{array}\right.$$

$$\tag{3.5} \left\{\begin{array}{rl} r_4 &= R_m(\tfrac{1}{6}L_mP_3 + L_mq_3), \\ y_{j+1} &= z_4 = z_3 + r_4\,, \\ L_mq_{4,j+1} &= 3\,[L_mr_4 - (\tfrac{1}{6}L_mP_3 + L_mq_3)]. \end{array}\right.$$

Remark 1. Regarding all operations in (3.1)–(3.5) as exact, we can replace R_m by 10^{-m} and L_m by 10^m. A straightforward calculation then shows that (3.1)–(3.5) is equivalent to (2.1)–(2.5).

Remark 2. The quantities, r_i, are redundant if all operations in (3.1)–(3.5) are considered to be exact. They play a significant role only when digital numbers and pseudo-operations are introduced.

Remark 3. The r_i require only one additional storage register rather than n. The order of computation should be as follows. First, compute together the components of r_i, z_i, and L_mq_i, as indicated by the inner brackets. Then the components of L_mP_i can be computed. Since the computation of a component of z_i and L_mq_i requires only the corresponding component of r_i, and since r_i is not used after the ith stage, one storage suffices for all components of all r_i.

Remark 4. It is obvious that the quantity, L_mq_{4j}, is always zero if exact operations are used. For pseudo-operations this is not the case. However, L_mq_{40} can always be taken to be zero to start the computation.

In practice, the finite-precision procedure, (3.1)–(3.5), would be carried out with digital numbers and pseudo-operations. The operation "+" would be executed as pseudo-addition, and both R_m and L_m would be executed as shift operations rather than as multiplications. To analyze the rounding error, it is convenient to rewrite (3.1)–(3.5) in a mixed form, $\overline{(3.1)}$–$\overline{(3.5)}$, involving digital numbers, exact operations, and error terms. It is to be understood that the digital numbers are thereby treated as real decimal numbers having zeros in all positions beyond position s. The effect of pseudo-operations is shown by the presence of a single error term denoted by an expression of the form, $e(u)$.

$$\overline{(3.1)} \qquad \begin{cases} \bar{z}_0 = \bar{y}_j \\ \overline{L_m q_0} = \overline{L_m q_{4j}}\,, \\ \overline{L_m P_0} = (L_m \bar{h}) f(\bar{z}_0) + e(L_m P_0), \end{cases}$$

$$\overline{(3.2)} \qquad \begin{cases} \bar{r}_1 = R_m(\tfrac{1}{2}\overline{L_m q_0} - \overline{L_m P_0} + e(r_1), \\ \bar{z}_1 = \bar{z}_0 + \bar{r}_1\,, \\ \overline{L_m q_1} = 3L_m \bar{r}_1 - (\tfrac{1}{2}\overline{L_m P_0} - \overline{L_m q_0}) + e(L_m q_1), \\ \overline{L_m P_1} = (L_m \bar{h}) f(\bar{z}_1) + e(L_m P_1), \end{cases}$$

$$\overline{(3.3)} \qquad \begin{cases} \bar{r}_2 = R_m(\tfrac{1}{2}(\overline{L_m P_1} - \overline{L_m q_1})) + e(r_2), \\ \bar{z}_2 = \bar{z}_1 + \bar{r}_2\,, \\ \overline{L_m q_2} = -L_m \bar{r}_2 - \tfrac{1}{3}\overline{L_m q_1} + \tfrac{1}{2}\overline{L_m P_1} + e(L_m q_2), \\ \overline{L_m P_2} = (L_m \bar{h}) f(\bar{z}_2) - \tfrac{1}{2}\overline{L_m P_1} + e(L_m P_2), \end{cases}$$

$$\overline{(3.4)} \qquad \begin{cases} \bar{r}_3 = R_m(\overline{L_m P_2}) + e(r_3), \\ \bar{z}_3 = \bar{z}_2 + \bar{r}_3\,, \\ \overline{L_m q_3} = -L_m \bar{r}_3 + \overline{L_m q_2}\,, \\ \overline{L_m P_3} = (L_m \bar{h}) f(\bar{z}_3) + 2\overline{L_m P_2} + e(L_m P_3), \end{cases}$$

$$\overline{(3.5)} \qquad \begin{cases} \bar{r}_4 = R_m(\tfrac{1}{6}\overline{L_m P_3} + \overline{L_m q_3}) + e(r_4), \\ \bar{y}_{j+1} = \bar{z}_4 = \bar{z}_3 + \bar{r}_4\,, \\ \overline{L_m q_{4,j+1}} = 3[L_m \bar{r}_4 - (\tfrac{1}{6}\overline{L_m P_3} + \overline{L_m q_3})] + e(L_m q_4). \end{cases}$$

Remark 5. It is seen that $e(z_i) = 0$ for $i = 0, 1, \cdots, 4$ because the pseudo-operation of addition gives the same result as the exact operation. This is true because of our assumption that in all digital numbers the decimal point is in a fixed position. If "floating-point" numbers are used, the pseudo-operation of addition can introduce a rounding error. We shall discuss this in the next section.

We are now in a position to estimate the rounding error in (3.1)–(3.5). After some preliminaries, we shall formulate the results as Theorem 2 and its corollary.

For any quantity, u, we define $\epsilon(u) = \bar{u} - u$; i.e. $\epsilon(u)$ is the total rounding error in u. We are interested in $\epsilon(y_j)$. However, it will turn out that the quantity,

$$\bar{y}_j^* = \bar{y}_j - \frac{R_m}{3}\,\overline{L_m\, q_{4j}}\,,$$

is a better approximation to y_j than is $\bar{y}_j$. Thus, we shall consider $\epsilon(y_j^*)$ instead, where

$$y_j^* = y_j - \frac{R_m}{3}\, L_m\, q_{4j} = y_j - \frac{1}{3}\, q_{4j}\,.$$

By remark 4, $q_{4j} = 0$, so that $y_j^* = y_j$. Hence,

$$y_j = \bar{y}_j^* - \epsilon(y_j^*).$$

We note that

$$\epsilon(y_j^*) = \epsilon(y_j) - \frac{R_m}{3}\,\epsilon(L_m q_{4j}).$$

It is convenient to deal with the norm of a vector, u, which we define as

$$\| u \| = \max_k | u_k |,$$

where u_k are the components of u. For a matrix, A, with elements, a_{ik}, we define

$$\| A \| = \max_i \{ \sum_k | a_{ik} | \}.$$

In particular, we shall be concerned with matrices for which $a_{ik} = \partial f_i / \partial y_k$, where the partial derivatives are evaluated at different points for each i and k. A matrix of this type will be denoted by the symbol, "J."

THEOREM 2. *For any of the quantities, u, computed in* ($\overline{3.1}$)–($\overline{3.5}$), *let the error term, $e(u)$, be subject to the condition,*

(i) $$\| e(u) \| \leq \frac{M}{2}\, 10^{-s}.$$

Let the bounds on the partial derivatives, $\partial f_i / \partial y_k$, be such that for any matrix, J,

(ii) $$\| J \| \leq L.$$

Let $h = 10^{-m}$, $0 < m < s$. Then the total rounding error in y_j^* incurred in one integration step is not greater than $2M10^{-s-m}$ in absolute value.

Proof. From (3.2) and ($\overline{3.2}$) we obtain

$$\epsilon(z_1) = \epsilon(z_0) + R_m(\tfrac{1}{2}\epsilon(L_m P_0) - \epsilon(L_m q_0)) + e(r_1).$$

From (3.1), ($\overline{3.1}$), if we assume that $h = \bar{h}$, we have

$$\epsilon(L_m P_0) = L_m(h)(f(\bar{y}_j) - f(y_j)) + e(L_m P_0).$$

Now, for each component, f_i, $i = 0, 1, \cdots, n$, we have

$$f_i(\bar{y}_j) - f_i(y_j) = \sum_{k=0}^{n} \frac{\partial f_i}{\partial y_{jk}}\, \epsilon(y_{jk}),$$

or, in matrix notation,

$$f(\bar{y}_j) - f(y_j) = J\epsilon(y_j).$$

This gives

$$\epsilon(L_m P_0) = L_m(h) J \epsilon(y_j) + e(L_m P_0),$$

(3.6)

$$\epsilon(z_1) = \epsilon(y_j) + \frac{h}{2} J \epsilon(y_j) - R_m\, \epsilon(L_m q_{4j}) + \frac{R_m}{2}\, e(L_m P_0) + e(r_1).$$

Proceeding in this way, we obtain

(3.7) $$\epsilon(z_2) = \left(I + \frac{h}{2} J + \frac{h^2}{4} J^2\right)\epsilon(y_j) + e(r_2) - \frac{1}{2}\, e(r_1) + z$$

where

$$z = \frac{hJ}{2} e(r_1) + \frac{R_m}{2} e(L_m P_1) - \frac{R_m}{2} e(L_m q_1) - \frac{hR_m}{2} J[\epsilon(L_m q_{4j}) - e(L_m P_0)],$$

$$\begin{aligned} \epsilon(z_3) = \left(I + hJ + \frac{h^2}{2} J^2 + \frac{h^3}{4} J^3\right) \epsilon(y_j) - \frac{1}{2} e(r_1) + e(r_2) + e(r_3) \\ + R_m e(L_m P_2) - \frac{R_m}{2} e(L_m q_1) + hJe(r_2) - \frac{h}{2} Je(r_1) + hJz, \end{aligned} \tag{3.8}$$

$$\epsilon(y_{j+1}) = \epsilon(y_j) - \frac{R_m}{3} \epsilon(L_m q_{4j}) + hJv + \frac{R_m}{6} W + e(r_4), \tag{3.9}$$

where

$$v = \tfrac{1}{6}(\epsilon(y_j) + 2\epsilon(z_1) + 2\epsilon(z_2) + \epsilon(z_3)),$$

and

$$W = e(L_m P_0) + 2e(L_m P_1) + 2e(L_m P_2) + e(L_m P_3) + 6e(L_m q_2) - 2R_m e(L_m q_1).$$

Using (3.6)–(3.8), we obtain

$$\begin{aligned} v = \left(I + \frac{h}{2} J + \frac{h^2}{6} J^2 + \frac{h^3}{24} J^3\right) \epsilon(y_j) - \frac{R_m}{3} \epsilon(L_m q_{4j}) \\ + \frac{1}{12} e(r_1) + \frac{1}{2} e(r_2) + \frac{1}{6} e(r_3) + \mu_1 + \mu_2, \end{aligned} \tag{3.10}$$

where

$$\mu_1 = \frac{R_m}{6} (e(L_m P_0) + e(L_m P_1) + e(L_m P_2)) - \frac{R_m}{4} e(L_m q_1),$$

$$\mu_2 = \frac{z}{3} + \frac{h}{6} J(e(r_2) - \frac{1}{2} e(r_1) + z).$$

From $(\overline{3.5})$ and the fact that $q_{4j} = 0$, we obtain

$$\epsilon(L_m q_{4,j+1}) = 3L_m e(r_4) + e(L_m q_4). \tag{3.11}$$

If we multiply (3.11) by $R_m/3$ and subtract from (3.9), we obtain

$$\epsilon(y_{j+1}^*) = \epsilon(y_j^*) + hJv + \frac{R_m}{6} W - \frac{R_m}{3} e(L_m q_4). \tag{3.12}$$

Applying the properties of the norm and using conditions (i) and (ii), we get

$$\begin{aligned} \| z \| &\leqq h \| J \| \cdot \| e(r_1) \| + \frac{R_m}{2} \| e(L_m P_1) \| + \frac{R_m}{2} \| e(L_m q_1) \| \\ &+ \frac{hR_m}{2} \| J \| \cdot \| e(L_m P_0 \| + \frac{hR_m}{2} \| J \| \cdot \| \epsilon(L_m q_{4j}) \|, \\ &\leqq \frac{1}{2} (hL + R_m) 10^{-s} M + \frac{LhR_m}{4} 10^{-s} M + \frac{hR_m}{2} \| \epsilon(L_m q_{4j}) \| L. \end{aligned}$$

Continuing in this way, we have

$$\| \mu_2 \| \leqq \frac{1}{3} \| z \| + \frac{hL}{8} 10^{-s}M + h \| z \| L,$$

$$\| \mu_1 \| \leqq \tfrac{3}{8} R_m 10^{-s}M,$$

$$\| W \| \leqq 6(10^{-s})M + R_m 10^{-s}M,$$

$$\| v \| \leqq \left(1 + \frac{hL}{2} + \frac{h^2L^2}{6} + \frac{h^3L^3}{24}\right) \| \epsilon(y_j) \| + \frac{R_m}{3} \| \epsilon(L_m\, q_{4j}) \| + \tfrac{3}{8} 10^{-s}M + \| \mu_1 \| + \| \mu_2 \|,$$

$$\| \epsilon(y^*_{j+1}) \| \leqq \| \epsilon(y_j^*) \| + Lh \| v \| + \frac{R_m}{6} \| W \| + \frac{R_m}{3} \| e(L_m\, q_4) \|,$$

$$\begin{aligned} \| \epsilon(y^*_{j+1}) \| \leqq \| \epsilon(y_j^*) \| &+ \left(hL + \frac{h^2L^2}{2} + \frac{h^3L^3}{6} + \frac{h^4L^4}{24}\right) \| \epsilon(y_j) \| \\ &+ hR_m \left(\frac{L}{3} + \frac{hL^2}{6} + \frac{h^2L^3}{2}\right) \| \epsilon(L_m\, q_{4j}) \|, \\ &+ \left(\frac{4}{3} R_m + \frac{3}{8} h\right) 10^{-s}M + \left(\frac{R_m^{\,2}}{6} + \frac{13}{24} hR_m + \frac{7h^2}{24}\right) 10^{-s}M \\ &+ \left(\frac{h^3}{2} + \frac{7h^2R_m}{12}\right) 10^{-s}M + \left(\frac{h^3R_m}{4}\right) 10^{-s}M. \end{aligned} \tag{3.13}$$

Now, to estimate the rounding error incurred in *one* integration step, say from j to $j + 1$, we assume that all quantities obtained at the jth step are exact. Thus, in (3.13) we set $\epsilon(y_j^*) = 0$, $\epsilon(L_m q_{4j}) = 0$, and $\epsilon(y_j) = 0$. Denoting the one-step rounding error by $\epsilon_1(y^*_{j+1})$, we have

$$\begin{aligned} \| \epsilon_1(y^*_{j+1}) \| \leqq \left(\frac{4}{3} R_m + \frac{3}{8} h\right) 10^{-s}M &+ \left(\frac{R_m^{\,2}}{6} + \frac{13}{24} hR_m + \frac{7h^2}{24}\right) 10^{-s}M \\ &+ \left(\frac{h^3}{2} + \frac{7}{12} h^2R_m\right) 10^{-s}M + \frac{h^3R_m}{4} 10^{-s}M. \end{aligned} \tag{3.14}$$

Since $h = 10^{-m}$, we can take $R_m = h$ and get

$$\| \epsilon_1(y^*_{j+1}) \| \leqq \left[\frac{41}{24} 10^{-s-m} + 10^{-s-2m} + \frac{13}{12} 10^{-s-3m} + \frac{1}{4} 10^{-s-4m}\right] M, \tag{3.15}$$

which proves the theorem.

COROLLARY. *A bound for the accumulated rounding error, under the hypotheses of Theorem 2, is given by*

$$\| \epsilon(y_j^*) \| \leqq \| \epsilon(y_0^*) \| e^{hjL} + (1 - e^{hjL})\left(\frac{f(h)}{1 - e^{hL}}\right) 10^{-s}M, \tag{3.16}$$

where $f(h) = \dfrac{h}{48}\left(130 + 160h + 100h^2 + 27h^3 + \dfrac{h^4}{3}\right)$.

Proof. From the definition of y_j^* we obtain

$$\| \epsilon(y_j) \| \leqq \| \epsilon(y_j^*) \| + \frac{R_m}{3} \| \epsilon(L_m\, q_{4j}) \| .$$

From (3.11), we have

$$R_m \| \epsilon(L_m\, q_{4j}) \| \leqq \frac{3}{2}(10^{-s})M + \frac{R_m}{2} 10^{-s}M,$$

Using (3.13), we get

$$\| \epsilon(y_{j+1}^*) \| \leqq e^{hL} \| \epsilon(y_j^*) \| + (f_1(h) + f_2(R_m, h))10^{-s}M, \tag{3.17}$$

where

$$f_1(h) = \frac{h}{48}(66 + 110h + 64h^2 + h^3),$$

$$f_2(h) = \frac{R_m}{48}\left(64 + 8R_m + 42h + 36h^2 + 26h^3 + \frac{h^4}{3}\right).$$

Setting $R_m = h$, and solving the difference equation corresponding to (3.17), we obtain (3.16).

Remark 6. Theorem 2 gives an upper bound on the one-step rounding error. A somewhat better result can be obtained from a statistical estimate of this error, if one is willing to make certain assumptions. If we assume (1) that the components of the errors, $e(r_i)$ and $e(LP_i)$ are independent and have a uniform distribution between $-10^{-s}/2$ and $10^{-s}/2$, and (2) that the bias which would be introduced in $e(Lq_1)$ and $e(Lq_2)$ by the coefficient, $\frac{1}{2}$, is eliminated by 'rounding up' Lq_1 and 'rounding down' Lq_2, then a direct computation with (3.12) yields as the approximate standard deviation of a component of $\epsilon(y_j^*)$,

$$\sigma_i \doteq \frac{1}{6}\left[\frac{10}{3}\left(R_m^2 + \frac{h^2}{4}\sum_k (\partial f_i/\partial y_k)^2\right)\right]^{\frac{1}{2}} 10^{-s}. \tag{3.18}$$

This is approximately the standard deviation of an error which is uniformly distributed between $\pm R_m 10^{-s}/2$, so that the accuracy is the same as would be obtained with $s + m$ digits.

Remark 7. As an example, we follow Gill [1] and integrate $y' = y$ from $t = 0$ to $t = 1$, with $h = 0.1$ and $s = 6$. The results are given in Table 1. The values in parentheses are those obtained by Gill's method [1]. For $t = 1$, after ten steps, we should have the value of $e/10$. If we use $y - q/3$ for this value, we obtain 0.27182810, which is in error by -8×10^{-8}. (Note that in computing q_1 the result of multiplying by $\frac{1}{2}$ was rounded up, while in the computation of q_2, it was rounded down.)

Remark 8. It is of interest to compare the accumulated error of the above example with the bounds given by (3.16) of the corollary and by statistical estimates. Since $\epsilon(y_0^*) = 0$ in the example, and $t = hj = 1$, (3.16) becomes

$$| \epsilon(y_j^*) | \leqq (e - 1)\frac{f(h)}{(e^h - 1)} \times 10^6 \leqq 1.72\frac{f(h) \times 10^{-6}}{(e^h - 1)} .$$

TABLE 1

Comparison of Gill's Method with the Modified Runge-Kutta Method

t	*Stage*	*r*	*z*	*lq*	*LP*
0.0	0	5 000	100 000	0	100 000
	1	250	105 000	100 000	105 000
	2	5 275	105 250	16 667	52 750
	3	110 525	110 525	− 36 083	216 025
0.1	4	− 8	(110 517) 110 517	(− 3) − 3	110 517
	1	5 526	116 043	110 518	116 043
	2	276	116 319	18 422	58 297
	3	5 830	122 149	− 39 878	238 744
0.2	4	− 9	(122 140) 122 140	(− 8) − 9	122 140
	1	6 108	128 248	122 161	128 248
	2	304	128 552	20 364	64 428
	3	6 443	134 995	− 44 066	263 851
0.3	4	− 9	(134 986) 134 986	(+ 4) + 3	134 986
	1	6 749	141 735	134 980	141 735
	2	338	142 073	22 494	71 206
	3	7 121	149 194	− 48 716	291 606
0.4	4	− 12	(149 182) 149 182	(− 14) − 15	149 182
	1	7 461	156 643	149 224	156 643
	2	371	157 014	24 870	78 693
	3	7 869	164 883	− 53 820	322 269
0.5	4	− 11	(164 872) 164 872	(− 4) − 3	164 872
	1	8 244	173 116	164 881	173 116
	2	412	173 528	27 478	86 970
	3	8 697	182 225	− 59 492	356 165
0.6	4	− 13	(182 212) 182 212	(+ 3) + 3	182 212
	1	9 110	191 322	182 197	191 322
	2	456	191 778	30 369	96 117
	3	9 612	201 390	− 65 751	393 624
0.7	4	− 15	(201 375) 201 375	(− 9) − 9	201 375
	1	10 070	211 445	201 403	211 445
	2	502	211 947	33 568	106 225
	3	10 623	222 570	− 72 662	435 020
0.8	4	− 16	222 554 222 554	(− 3) − 3	222 554
	1	11 128	233 682	222 560	233 682
	2	556	234 238	37 094	117 397
	3	11 740	245 978	− 80 306	480 772
0.9	4	− 18	(245 960) 245 960	(− 9) − 9	480 772
	1	12 299	258 259	245 981	258 259
	2	614	258 873	040 995	129 744
	3	12 974	271 847	− 88 745	531 335
1.0	4	− 19	(271 828) 271 828	(− 4) − 3	

Now,

$$g(h) = \frac{f(h)}{(e^h - 1)} \doteq \frac{130 + 160h + 100h^2 + 27h^3}{48(1 + h/2 + h^2/6 + h^3/24)},$$

and $g(.1) \doteq 2.92$. Hence,

$$| \epsilon(y_j^*) | \leqq 4.98 \times 10^{-6}.$$

To obtain a statistical estimate, we might assume that accumulated error is the sum of the one-step errors and that these errors are independent. Using (3.18), the standard deviation for one step is about 3.4×10^{-8}. The standard deviation after ten steps is $\sqrt{10}$ times this, or about 1.1×10^{-7}.

It is of interest to compare the above results with those obtained from the classical Runge-Kutta method, (1.5)–(1.9). These are tabulated below.

t	y
0.0	.100 000
0.1	110 517
0.2	122 140
0.3	134 986
0.4	149 183
0.5	164 873
0.6	182 213
0.7	201 377
0.8	222 556
0.9	245 963
1.0	271 831

4. Floating-point Arithmetic. Since many modern automatic computers provide "floating-point" operations, and since the finite-precision modification must be applied in a slightly different way when floating-point numbers are used, it seems worthwhile to devote a short section to this subject.

Let us begin by establishing certain conventions. A "digital number in normal floating-point form" consists of two parts, a modulus and an exponent. The modulus is an aggregate of s decimal digits, the decimal point being placed at the extreme left and the digit in position one being non-zero. The exponent consists of two digits and represents the power of ten which multiplies the modulus. An algebraic sign is associated with each modulus and exponent. Thus, the fixed-point number, .00113, would be written as +.113–02 in normal floating form, and −11.3 would be written as −.113 + 02. In floating-point arithmetic some of the shift operations of formulas (3.1)–(3.5) will be carried out automatically by the positioning which must take place in the process of addition or subtraction. The rounding error will then be governed by the magnitude of h and the relative magnitudes of y and y'. If $hy_j' < y_j$, then Theorem 2 will apply to the procedure (4.1)–(4.5) below, it being understood that all errors must be considered as relative errors. If $hy_j' > y_j$ for some j, the theorem no longer holds. Nevertheless, over an interval, there should be a preponderance of points for which $hy_j' < y_j$, so that (4.1)–(4.5) should reduce the overall rounding error.

To explain the meaning of the symbols L and R in (4.1)–(4.5), we must first point out that in automatic computers, the exponent of a floating-point number is placed to the left of the modulus. Thus, a shift right m places will not affect the exponent if $m < s$. Now, in the computation of r_i, the exponent, μ, of the quantity in square brackets is compared with the exponent, ρ, of z_{i-1}. If $\mu < \rho$, then $R^{(i)} = R_{\rho-\mu}$ and $L^{(i)} = L_{\rho-\mu}$. If $\mu \geqq \rho$, then $R^{(i)} = R_0$ and $L^{(i)} = L_0$. With this interpretation of the shift operations, the finite-precision modification for floating-point arithmetic is as follows:

$$(4.1)\quad \begin{cases} z_0 = y_j, \\ q_0 = q_{4j}, \\ P_0 = hf(z_0), \end{cases}$$

$$(4.2)\quad \begin{cases} \begin{cases} r_1 = L^{(1)}R^{(1)}[\tfrac{1}{2}P_0 - q_0], \\ z_1 = z_0 + r_1, \\ q_1 = 3r_1 - (\tfrac{1}{2}P_0 - q_0), \end{cases} \\ P_1 = hf(z_1), \end{cases}$$

$$(4.3)\quad \begin{cases} \begin{cases} r_2 = L^{(2)}R^{(2)}[\tfrac{1}{2}(P_1 - q_1)], \\ z_2 = z_1 + r_2, \\ q_2 = -r_2 - \tfrac{1}{3}q_1 + \tfrac{1}{2}P_1, \end{cases} \\ P_2 = hf(z_2) - \tfrac{1}{2}P_1, \end{cases}$$

$$(4.4)\quad \begin{cases} \begin{cases} r_3 = L^{(3)}R^{(3)}[P_2], \\ z_3 = z_2 + r_3, \\ q_3 = -r_3 + q_2, \end{cases} \\ P_3 = hf(z_3) + 2P_2, \end{cases}$$

$$(4.5)\quad \begin{cases} r_4 = L^{(4)}R^{(4)}[\tfrac{1}{6}P_3 + q_3], \\ y_{j+1} = z_4 = z_3 + r_4, \\ q_{4,j+1} = 3[r_4 - (\tfrac{1}{6}P_3 + q_3)]. \end{cases}$$

Computation and Data Reduction Center
Space Technology Laboratories, Inc.
Los Angeles 45, California

1. S. Gill, "A process for the step-by-step integration of differential equations in an automatic digital computing machine," *Proc. Cambridge Philos. Soc.*, v. 47, pt. 1, p. 96–108.

2. J. von Neumann, & H. H. Goldstine, "Numerical inverting of high order matrices," *Bull. Amer. Math. Soc.*, v. 53, n. 11, November 1947.

A Statistical Study of Randomness Among the First 10,000 Digits of π

By R. K. Pathria

1. Introduction. In connection with the application of Monte Carlo methods to various problems in mathematical physics and the drawing of random samples in statistics there arose a demand for the so-called random digits. As a result of the rapid progress made in these fields of investigation this demand has increased considerably during recent years. Consequently, a number of standard sets of such digits have been produced and are being put to frequent use by workers engaged in these fields [1]-[4].

At this stage it appears worthwhile to investigate, as has recently been suggested by the author [5], the extent to which one can utilize the digits appearing in the decimal development of the various constants of mathematical analysis, such as e, π, etc., for the purposes mentioned above. It is obvious that such a suggestion would have been hardly of any practical interest if it had been made at a time when the values of these constants were not yet available to a reasonably large number of decimal places. However, certain computations of this type have been carried out during recent years and in the near future they are to be extended to the point where they will surely provide sets of digits as large as the existing ones.*

Obviously, the question of randomness of the digits to be studied here cannot be decided on *a priori* grounds. One has to subject them to various tests and obtain internal evidence for their randomness before they can be declared fit for practical use. It appears worthwhile to mention here that apart from the specific purposes indicated above, a study of this type is fascinating also because of its intrinsic interest. It was apparently for this latter reason that Reitwiesner [6], at the suggestion of von Neumann, computed the values of π and e to more than 2,000 decimal places and Metropolis, Reitwiesner and von Neumann [7] carried out a statistical treatment thereof by studying the frequency distribution of the various digits. This study was extended to about 3,000 decimal places by Gruenberger [8] in the case of e and by Nicholson and Jeenel [9] in the case of π.

In the present paper a report is given of the results obtained by applying the four classical tests of Kendall and Smith (the frequency test, the serial test, the poker test, and the gap test) [10] and a fifth one due to Yule (the five-digit sum

Received February 27, 1961.

* Dr. D. B. Gillies of the Digital Computer Laboratory, University of Illinois, has kindly informed the author that in a year or so they will probably compute one million digits of e. At present their computation extends to 60,000 decimal places. A statistical study of these digits is also being carried out by the author and will be reported shortly.

Very recently Dr. Shanks and Dr. Wrench have made an IBM 7090 calculation of both π and e to 100,000 decimal places. The frequency distribution of the decimal digits of both the constants has also been computed. The author is highly grateful to Dr. Wrench for illuminating communications on this subject.

test) [11] to the first 10,000 digits of π.† The value investigated here is the one computed by Genuys [13] using the formula

$$\pi = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)} \left\{ \frac{16}{5^{2k+1}} - \frac{4}{239^{2k+1}} \right\},$$

and computing its terms on the IBM 704. The present analysis has been carried out mostly in blocks of 1,000 digits each, with a view to discover 'patches,' if any, that suffer from lack of local randomness. Of course, blocks which are found patchy are not suitable for drawing a random sample when used by themselves. They have to be suitably diluted by combining them with some of the neighboring blocks in order to obtain larger ones which could safely be employed in a statistical investigation.

In comparing the actual frequencies with expectations the χ^2 test has mostly been employed; the rejection levels, following Kendall and Smith [2], have been kept at 1 and 99 per cent.

2. The Frequency Tests. The 10,000 digits of $\pi - 3$ have been divided into ten consecutive blocks of 1,000 digits each and the frequencies f_i with which the various digits $i(= 0, 1, \cdots, 9)$ appear in these blocks have been recorded. These frequencies, along with the respective values of the statistic χ^2 and the corresponding probabilities P for nine degrees of freedom, are given in Table 1. It is only in the case of the third and the ninth blocks that the value of P is found to be significant; in the former case the deviations from the expected frequencies are too high, while in the latter they are too low.

Taking the table as a whole, of the 100 frequencies recorded 34 deviate from the expected value of 100 by more than the standard deviation $\sigma(= \sqrt{90} = 9.487)$ and 6 by more than 2σ. These figures compare well with the corresponding ones, namely, 31.73 and 4.55 per cent, for a normal distribution. Further, in the case of total frequencies the χ^2 value (9.318 for 9 d.f.) may be partitioned into three components, with the following obviously satisfactory results:

Classification	χ^2	d.f.	P
Odd versus even digits	0.360	1	~55%
Within groups of odd digits	4.358	4	~35%
Within groups of even digits	4.602	4	~35%

3. The Serial Tests. These tests are employed with a view of looking for any evidence of serial association among the digits under study. The relevant test here consists in classifying the digit pairs (ij) with respect to the members i and j comprising a pair and comparing the frequencies thus obtained with expectations. We have tabulated the frequencies for the 10,000 overlapping pairs, formed by the first

† Gruenberger [12] has shown how the tests given by Kendall and Smith can be applied to any set of digits, punched on IBM cards, mechanically and without regard to the order of the digits on the cards, using standard IBM equipment. In the absence of such a facility, however, the author has made the various tabulations by hand and has satisfied himself about their correctness by applying suitable cross-checks.

TABLE 1
Frequency Distribution Among the First 10,000 Digits of $\pi - 3$

Block \ Digit	0	1	2	3	4	5	6	7	8	9	χ^2	$P(\%)$
1	93	116	103	102	93	97	94	95	101	106	4.74	~85
2	89	96	104	86	102	108	106	102	101	106	4.94	~85
3	77	97	96	77	123	110	102	90	108	120	22.80	<1
4	103	120	105	103	87	102	96	90	95	99	7.58	~60
5	104	103	88	91	103	108	115	111	87	90	9.38	~40
6	91	94	98	113	105	97	106	118	90	88	9.28	~40
7	100	107	98	114	89	108	89	88	98	109	7.84	~55
8	97	100	119	95	107	104	108	92	84	94	8.80	~45
9	101	103	100	103	101	99	98	97	90	108	1.98	>99
10	113	90	110	90	102	113	107	87	94	94	9.32	~40
(1–10)*	968	1026	1021	974	1012	1046	1021	970	948	1014	9.318	~40

* The cumulative frequencies obtaining in this row are in complete agreement with the ones given by Dr. Wrench (private communication). See also J. W. Wrench, Jr., "The evolution of extended decimal approximations to π," *The Math. Teacher*, v. 53, 1960, p. 644–650; v. 55, 1962, p. 129–130.

10,001 digits of π, in Table 2. The following relations exist among these frequencies:

$$\sum_{i,j} f_{ij} = N$$

and

$$\sum_{l} f_{lm} = \sum_{n} f_{mn} + \epsilon_m ,$$

where $N = 10{,}000$ and ϵ_m which represents the "end effects" is equal to zero if the digit m appears either at both the ends of the set or at none; it is equal to -1 if the set opens with m and $+1$ if the set closes with m. In the case under study, we have $\epsilon_3 = -1$ and $\epsilon_8 = +1$. As a final check on the entries in this table, one verifies that the sum $\sum_{i,j} f_{ij}(i - j)$, which should obviously be equal to the difference between the first and the last digits of the set, is really equal to -5.

Now, the overall expectation m_{ij} of f_{ij} is, for each of the pairs, equal to Np^2, where p is the probability of occurrence of a particular digit. The variance of f_{ij} is, however, given by

$$\sigma_{ij}^2 = Np^2(1 + 2p\delta_{ij} - 3p^2),$$

where δ_{ij} is the Kronecker delta. Thus, whereas the expectation for each of the hundred elements of the array is 100 on the basis of perfect randomness, the standard deviation for the diagonal elements is 10.82 and that for the non-diagonal ones is 9.85. The observed values of the root-mean-square deviation are 9.76 and 8.78, respectively. Comparing the differences with the standard error in the dispersion one finds that none of these values is significant.

Several essentially equivalent values of χ^2 have been computed from Table 2. First, assuming all the hundred types of pairs to be equally likely (expected value of 100 for each cell), a χ^2 of 78.84 is obtained which, for 90 d.f., is at about 80 per cent probability level. Second, given the row sums and assuming the ten digits to be equally likely to follow (e.g., expected value of 96.8 for each of the cells in the first row), a χ^2 of 69.39 is obtained which, for 90 d.f., is at about 95 per cent prob-

TABLE 2

Frequency Distribution Among the First 10,000 Overlapping Pairs (ij) of π(= 3.14 ··· 78)

j / i	0	1	2	3	4	5	6	7	8	9	Total
0	85	103	98	103	98	89	101	93	83	115	968
1	99	99	103	102	121	95	106	90	98	113	1026
2	101	115	110	99	82	118	100	101	100	95	1021
3	102	92	86	94	114	100	90	102	97	98	975
4	95	100	100	89	102	110	103	108	101	104	1012
5	92	117	110	96	108	96	115	107	96	109	1046
6	107	95	117	97	101	124	91	101	90	98	1021
7	89	105	99	91	92	101	95	97	103	98	970
8	86	97	99	93	96	106	114	83	80	93	947
9	112	103	99	110	98	107	106	88	100	91	1014
Total	968	1026	1021	974	1012	1046	1021	970	948	1014	10000

ability level. Third, assuming the expectation of a particular cell to be one-tenth of the corresponding column sum, we get $\chi^2 = 69.26$ which, again for 90 d.f., gives $P \simeq 95$ per cent. Fourth, fitting all the expectations to both the row sum and the column sum, a value of 59.83 results which, for 81 d.f., is at about 96 per cent probability level. All these figures are obviously satisfactory.

Next, we have computed from Table 2 the value of the quantity $\overline{ij}$ whose theoretical expectation and standard deviation are given by

$$E(\overline{ij}) = (\bar{\imath})^2$$

and

$$\sigma(\overline{ij}) = \{\overline{\imath^2} - (\bar{\imath})^2\} \cdot N^{-1/2}.$$

The actual value of this quantity turns out to be 20.062 which deviates from the expectation by an amount -1.1 times the S.D. The probability of equal or greater divergence of either sign is about 27 per cent—a result that is not significant.

So far we have been discussing the question of serial association between the neighboring digits comprising the whole set of 10,000 digits. We shall now study the various blocks one by one and see if they are individually also locally random. For this purpose, we give below the results of the χ^2 test, carried out on the assumption of equal *a priori* probability for each of the hundred cells:

Block	1	2	3	4	5	6	7	8	9	10
χ^2	96.6	82.2	115.2	101.4	96.2	135.4	90.8	93.0	80.6	100.2
P(%)	30	71	4	20	31	0.1	46	40	75	22

The P value in the case of the sixth block is too low and leads to its rejection outright.[‡] The only other block for which the P value is rather low is the third one; this, however, has already failed to meet the frequency test.

[‡] It may be noted that this block passed the frequency test very well. The failure here is mainly due to an essentially non-random arrangement of the digits in the block. For instance, the pair (77) appears 28 times (including 2 triplets and 3 quartets). Such an extreme pattern is dangerous even if diluted by one of its neighboring blocks. It can only be made harmless by combining with many other blocks.

TABLE 3

Classified Distribution of the First 2,000 Poker Hands of π

Classes	Actual Frequencies in Blocks										Expected Values	Actual Frequencies in the Whole Set	Expected Values
	1	2	3	4	5	6	7	8	9	10			
Busts (abcde)	63	54	74	72	68	58	51	71	60	58	60.48	629	604.8
Pairs (aabcd)	97	100	89	88	93	108	98	90	105	110	100.80	978	1008.0
Two pairs (aabbc)	23	30	25	27	22	18	32	17	15	18	21.60	227	216.0
Threes (aaabc)	14	13	11	10	13	14	15	19	18	13	14.40	140	144.0
Full house (aaabb)	2	3	1	3	1	0	3	2	2	1	1.80	18	18.0
Fours (aaaab)	1	0	0	0	3	2	1	1	0	0	0.90	8	9.0
Fives (aaaaa)	0	0	0	0	0	0	0	0	0	0	0.02	0	0.2
Total	200	200	200	200	200	200	200	200	200	200	200.00	2000	2000.0
χ^2(3 d.f.) →	0.34	4.04	6.47	6.16	1.55	1.29	6.78	5.36	2.68	2.11		2.69 (5 d.f.)	
$P(\%)$ →	95	25	10	10	50	50	8	15	45	55		75	

4. The Poker Test. The 10,000 digits of π are printed in 2,000 hands of five digits each. Among these hands we have noted the frequency of occurrence of those hands whose digits, with respect to their values, are either (i) all different, or (ii) one pair and the other three different, or (iii) two pairs and one different, or (iv) one triplet and two different, or (v) one triplet and one pair, or (vi) one quartet and one different, or finally (vii) one quintet. As usual, the frequencies thus obtained are compared with expectations. The results are given in Table 3. None of the χ^2 values is found to be significant.

An interesting observation may, however, be made here. Since the deviations in the third and the fourth blocks are, on the whole, in the same direction, a grouping of these two consecutive blocks results in a P value of about 1.5 per cent which is pretty low, though not below the rejection level. If, however, the rejection level were at 5 per cent, as might be the case in a more serious applicatior of these digits, this combined sample of 2,000 digits would no longer be considered locally random. In that case it would be essential to combine this sample with one of a sufficiently large strength before one could employ its digits in an investigation.

5. The Gap Test. Next, a frequency count has been made of the lengths of the gaps between successive zeros of the set. This frequency distribution is compared with the expected one only for the whole set and not for the individual blocks, because the frequencies in the latter case are too small unless, of course, a very coarse grouping of the classes is adopted.

TABLE 4

Length Distribution Among the Gaps Between Successive Zeros of π

Length of the Gap	Actual Frequency	Expected Frequency	Length of the Gap	Actual Frequency	Expected Frequency
0	85	100.00	16	18	18.53
1	78	90.00	17	13	16.68
2	87	81.00	18	14	15.01
3	68	72.90	19	13	13.51
4	80	65.61	20	10	12.16
5	61	59.05	21	13	10.94
6	47	53.14	22	10	9.85
7	49	47.83	23	9	8.86
8	38	43.05	24	7	7.98
9	42	38.74	25–29	32[1]	29.39
10	29	34.87	30–34	9[2]	17.36
11	30	31.38	35–39	8[3]	10.24
12	25	28.24	40–49	13[4]	9.64
13	29	25.42	$\geqq 50$	11[5]	5.15
14	25	22.88			
15	14	20.59	Total	967	1000.00

[1] 6, 10, 3, 7 and 6, in order of increasing length.

[2] 3, 3, 0, 1 and 2, in order of increasing length.

[3] 1, 2, 3, 2 and 0, in order of increasing length.

[4] 2 each, of lengths 40, 41, 43, 44, 46 and 47; 1 of length 48.

[5] Actual lengths are 51, 51, 54, 55, 60, 62, 63, 65, 65, 65 and 67.

We have 968 zeros in our set and hence 967 gaps; their length distribution§, along with the one expected on the basis of perfect randomness, is given in Table 4. The χ^2 value for the grouping as indicated in the table is 28.06 which, for 30 degrees of freedom, gives $P \simeq 55$ per cent—a result that is not significant.

The mean length of a gap (excluding those of length zero) is found to be 10.190 which deviates from the expected value of 10 by 0.601 times the standard deviation (viz., 0.316). The result is obviously satisfactory.

6. The Five-Digit Sum Test. This test, as applied here, consists in taking the sum of the five digits comprising a (poker) hand as the variable, denoted by the symbol i, say, and comparing its distribution over the various hands with the one expected theoretically. The latter may be obtained in an elegant manner as follows (refer to the alternative approach of Yule [11]).

If Ω_i denotes the number of ways in which the five digits of the hand can give a sum i, it will be enumerated by the generating function

$$\begin{aligned}\sum_{i=0}^{45} \Omega_i x^i &= [\sum_{k=0}^{9} x^k]^5 \\ &= (1 - x^{10})^5 \cdot (1 - x)^{-5} \\ &= f(x), \qquad \text{say.}\end{aligned}$$

It immediately follows that

$$\Omega_i = \sum_{r=0}^{5} (-1)^r \binom{5}{r} \binom{i - 10r + 4}{4}.$$

Moreover,

$$\sum_i \Omega_i = \underset{x \to 1}{Lt}\, f(x) = 10^5,$$

being the total number of ways in which a hand of five digits can be formed out of the digits of ten kinds. The probability p_i for the value i of the variable is then given by

$$p_i = \frac{\Omega_i}{\sum_i \Omega_i} = 10^{-5}\Omega_i,$$

which leads to the expected distribution. This is the same as the one given in Table I of reference [11]. The mean value of i is 22.5 and its standard deviation is

$$(41.25)^{1/2} = 6.4226.$$

The standard error of the mean of n observations is, therefore, equal to 6.4226 $n^{-1/2}$. Further, the standard error of the standard deviation turns out to be

$$(18.1/n)^{1/2} = 4.2544\, n^{-1/2}.$$

§ As a check, it has been verified that the total length of the 967 gaps, as tabulated here, is 8988 which, together with the 31 digits preceding the first zero and the 13 digits following the last one, makes 9,032—the number of non-zeros in the set.

TABLE 5

Five-Digit Sum Distribution Among the First 2,000 Hands of π

i	Expected Frequency in a Block of 400 Hands	Actual Frequencies in Blocks of 400 Hands Each					Actual Frequency in the Whole Set	Expected Frequency in the Whole Set
		I.	II	III	IV	V		
0	0.004	0	0	0	0	0	0	0.02
1	0.020	0	0	0	0	0	0	0.10
2	0.060	0	1	0	0	0	1	0.30
3	0.140	0	0	0	0	0	0	0.70
4	0.280	0	0	0	1	0	1	1.40
5	0.504	1	0	0	1	0	2	2.52
6	0.840	0	2	2	0	1	5	4.20
7	1.320	2	0	2	1	4	9	6.60
8	1.980	1	5	1	1	3	11	9.90
9	2.860	5	0	1	1	4	11	14.30
10	3.984	2	3	1	1	2	9	19.92
11	5.360	7	7	9	8	4	35	26.80
12	6.980	3	7	2	10	6	28	34.90
13	8.820	11	7	9	9	13	49	44.10
14	10.840	8	13	12	12	7	52	54.20
15	12.984	21	12	17	17	17	84	64.92
16	15.180	13	10	16	18	15	72	75.90
17	17.340	13	15	19	17	12	76	86.70
18	19.360	17	17	25	23	23	105	96.80
19	21.120	21	24	18	26	24	113	105.60
20	22.524	17	23	20	19	19	98	112.62
21	23.500	21	19	23	22	32	117	117.50
22	24.000	29	29	28	20	29	135	120.00
23	24.000	23	22	25	28	23	121	120.00
24	23.500	20	17	24	26	21	108	117.50
25	22.524	24	29	24	20	22	119	112.62
26	21.120	26	19	19	15	19	98	105.60
27	19.360	19	26	15	21	22	103	96.80
28	17.340	18	15	15	23	13	84	86.70
29	15.180	20	14	10	12	12	68	75.90
30	12.984	12	13	15	11	9	60	64.92
31	10.840	10	17	10	11	12	60	54.20
32	8.820	13	10	17	4	7	51	44.10
33	6.980	5	7	4	7	9	32	34.90
34	5.360	6	7	6	5	8	32	26.80
35	3.984	5	5	3	3	2	18	19.92
36	2.860	4	1	3	3	1	12	14.30
37	1.980	2	2	4	3	2	13	9.90
38	1.320	0	0	1	1	1	3	6.60
39	0.840	0	2	0	0	1	3	4.20
40	0.504	1	0	0	0	0	1	2.52
41	0.280	0	0	0	0	1	1	1.40
42	0.140	0	0	0	0	0	0	0.70
43	0.060	0	0	0	0	0	0	0.30
44	0.020	0	0	0	0	0	0	0.10
45	0.004	0	0	0	0	0	0	0.02
Total	400.000	400	400	400	400	400	2000	2000.00

TABLE 6

Mean Values of the Sum i, with Differences from Expectation, Etc.

Block	Mean Value	Difference from Expectation	Divided by Standard Error	Square of the Preceding Column
I	22.7450	+0.2450	+0.7629	0.5820
II	22.7550	+0.2550	+0.7941	0.6306
III	22.4625	−0.0375	−0.1168	0.0136
IV	22.0925	−0.4075	−1.2690	1.6104
V	22.1800	−0.3200	−0.9965	0.9930
The whole set	22.4470	−0.0530	−0.3690	

TABLE 7

Standard Deviations of the Sum i, with Differences from Expectation, Etc.

Block	Standard Deviation	Difference from Expectation	Divided by Standard Error	Square of the Preceding Column
I	6.3945	−0.0281	−0.1321	0.0174
II	6.4475	+0.0249	+0.1169	0.0137
III	6.2919	−0.1307	−0.6143	0.3773
IV	6.2241	−0.1986	−0.9334	0.8713
V	6.3716	−0.0510	−0.2398	0.0575
The whole set	6.3524	−0.0702	−0.7379	

The actual frequency distribution obtained from the 2,000 hands of the set is given in Table 5, where the results are also given for consecutive blocks of 400 hands each, i.e., comprising 2,000 digits each. The actual distribution is compared with the expected one through the mean values of the variable and its dispersions. In Table 6 we have listed for each of the five blocks, I to V, and for the whole set, the mean values and their deviations from the expectation in terms of the standard errors of the mean. None of the various deviations is found to be significant. In fact, the chance of equal or greater divergence, of either sign, in the case of the whole set is about 70 per cent. Moreover, even if we group together the last three blocks (each having a deviation of the same sign) the corresponding result comes out to be about 17 per cent. Still worse, if we take the last two blocks, for which the deviations are not only of the same sign but also of the greatest magnitude, the result is still about 11 per cent. Further, we note that the sum of the entries in the last column of the table is 3.83. Entering the χ^2 table with this value of χ^2 and 5 degrees of freedom, we find P to be about 60 per cent.

Finally, we study the standard deviations in the value of the variable as obtained from the frequencies tabulated above and compare them with the corresponding theoretical expectations. The relevant figures are given in Table 7. Expressing the deviations in terms of the standard errors of the standard deviation, we obtain results which do not exceed unity. Further, entering the χ^2 table with the sum of the squares of these numbers, namely, 1.34, and 5 degrees of freedom, we find that P lies between 90 and 95 per cent. For the whole set, the deviation of the actual standard deviation from the expected value is equal to -0.74 times the corresponding standard error; the chance of an equal or greater deviation of either sign is about 46 per cent.

Department of Physics
University of Delhi
Delhi-6, India

1. L. H. C. Tippett, *Random Sampling Numbers (41,600), Tracts for Computers, No. 15*, Cambridge University Press, 1927.

2. M. G. Kendall & B. B. Smith, *Random Sampling Numbers (100,000), Tracts for Computers, No. 24*, Cambridge University Press, 1939.

3. *Table of 105,000 x̄ Random Decimal Digits*, Interstate Commerce Commission, Bureau of Transport Economics and Statistics, Statement No. 4914, Washington, D. C., May 1949.

4. Rand Corporation. *A Million Random Digits with 100,000 Normal Deviates*. The Free Press, Glencoe, Illinois, 1955.

5. R. K. Pathria, "A statistical analysis of the first 2,500 decimal places of e and $1/e$," *Proc. Nat. Inst. Sci. India*, v. A27, 1961, p. 270–282. See also *Math. Gaz.*, London, v. 45, 1961, p. 142–143.

6. G. W. Reitwiesner, "An ENIAC determination of π and e to more than 2000 decimal places," *MTAC*, v. 4, 1950, p. 11–15.

7. N. C. Metropolis, G. Reitwiesner & J. von Neumann, "Statistical treatment of values of first 2000 decimal places of e and π calculated on the ENIAC," *MTAC*, v. 4, 1950, p. 109–111.

8. F. Gruenberger, "Further statistics on the digits of e," *MTAC*, v. 6, 1952, p. 123–124.

9. S. C. Nicholson & J. Jeenel," Some comments on a NORC computation of π," *MTAC*, v. 9, 1955, p. 162–164.

10. M. G. Kendall & B. B. Smith, "Randomness and random sampling numbers," *J. Roy. Statist. Soc.*, v. 101, 1938, p. 147–166.

11. G. U. Yule, "A test of Tippett's random sampling numbers," *J. Roy. Statist. Soc.*, v. 101, 1938, p. 167–172.

12. F. Gruenberger, "Tests on random digits," *MTAC*, v. 4, 1950, p. 244–245.

13. F. Genuys, "Dix mille decimales de π," *Chiffres*, v. 1, 1958, p. 17–22.

An Extended Table of Roots of $J_n'(x)Y_n'(\beta x) - J_n'(\beta x)Y_n'(x) = 0$

By John F. Bridge and Stanley W. Angrist

An eigen-equation that frequently occurs in mathematical physics involving an annular cavity is

$$J_n'(x)Y_n'(\beta x) - J_n'(\beta x)Y_n'(x) = 0 \tag{1}$$

where $J_n(x)$ and $Y_n(x)$ are respectively Bessel functions of the first and second kinds.

J. McMahon [1] gave an asymptotic expression for the roots to this equation, D. O. North [2] obtained a root smaller than the first root given by the asymptotic expression of McMahon, and R. Truell [2] developed a graphical method for obtaining this root.

H. B. Dwight [3] gave the first six roots of equation (1) for values of n from 1 to 3 and for various values of β from 1 to 4.

The purpose of this paper is to extend the range and accuracy of the roots to equation (1) and to determine the ranges of the solutions for which the asymptotic expression proposed by McMahon is sufficiently accurate.

The calculation of the roots was accomplished by trial and error substitution in the following equation:

$$F_n(x) = J_n'(x)Y_n'(\beta x) - J_n'(\beta x)Y_n'(x). \tag{2}$$

Starting with $x = 0.1$, $F_n(x)$ was calculated, increasing x in steps of 0.1 until $F_n(x)$ changed sign. A linear interpolation was then used to determine the approximate value of the root, then the Newton-Raphson iteration procedure was used until two successive approximations of the root value were within $\pm 10^{-6}$. The root thus obtained was compared with the root obtained with all four terms of J. McMahon's asymptotic expression (3)

$$x_n^{(s)} = \delta + \frac{p}{\delta} + \frac{q - p^2}{\delta^2} + \frac{r - 4pq + 2p^3}{\delta^5} + \cdots$$

$$\delta = \frac{(s - 1)}{\beta - 1}, \qquad p = \frac{m + 3}{8\beta}, \qquad m = 4n^2 \tag{3}$$

$$q = \frac{4(m^2 + 46m - 63)(\beta^3 - 1)}{3(8\beta)^3(\beta - 1)}$$

$$r = \frac{32(m^3 + 185m^2 - 2053m + 1899)(\beta^5 - 1)}{5(8\beta)(\beta - 1)}.$$

If the two values were within $\pm 2 \times 10^{-5}$, the asymptotic value was used for that root and all larger roots for the given value of n. If not, the procedure was continued until the next root was found and compared with its asymptotic value.

Received January 26, 1961. The research reported in this article was sponsored in part by the Rome Air Development Center, Air Research and Development Command.

TABLE 1
$\beta = 1.1$

s / n	1	2	3	4	5	6	7	8	9	10	11
1	0.952738	31.44125A	62.84451A	94.25622A	125.67004A	157.08470A	188.49978A	219.91511A	251.33059A	282.74616A	314.16181A
2	1.905471	31.48469A	62.86622A	94.27069A	125.68089A	157.09338A	188.50702A	219.92131A	251.33602A	282.75098A	314.16615A
3	2 858190	31.55696A	62.90238A	94.29480A	125.69898A	157.10785A	188.51907A	219.93165A	251.34506A	282.75902A	314.17338A
4	3.810883	31.65787A	62.95298A	94.32855A	125.72429A	157.12810A	188.53595A	219.94611A	251.35772A	282.77027A	314.18351A
5	4.763541	31.78713A	63.01796A	94.37192A	125.75683A	157.15414A	188.55765A	219.96471A	251.37399A	282.78474A	314.19653A
6	5.716165	31.94442A	63.09730A	94.42490A	125.79659A	157.18596A	188.58417A	219.98745A	251.39388A	282.80242A	314.21245A
7	6.668739	32.12933A	63.19094A	94.48747A	125.84356A	157.22355A	188.61550A	220.01430A	251.41739A	282.82332A	314.23125A
8	7.621261	32.34139A	63.29881A	94.55962A	125.89774A	157.26691A	188.65165A	220.04529A	251.44451A	282.84743A	314.25295A
9	8.573720	32.580055	63.42085A	94.64133A	125.95911A	157.31604A	188.69261A	220.08041A	251.47524A	282.87474A	314.27754A
10	9.526102	32.844769	63.56071A	94.73257A	126.02767A	157.37094A	188.73838A	220.11965A	251.50958A	282.90528A	314.30502A
11	10.478414	33.134922	63.70707A	94.83330A	126.10339A	157.43159A	188.78895A	220.16301A	251.54753A	282.93902A	314.33539A
12	11.430636	33.449843	63.87107A	94.94351A	126.18629A	157.49799A	188.84432A	220.21050A	251.58909A	282.97596A	314.36865A

TABLE 2
$\beta = 1.2$

s / n	1	2	3	4	5	6	7	8	9	10	11
1	0.910334	15.75443A	31.43914A	47.13936A	62.84346A	78.54910A	94.25552A	109.96238A	125.66951A	141.37683A	157.08428A
2	1.820564	15.83425A	31.47819A	47.16590A	62.86336A	78.56502A	94.26878A	109.97374A	125.67946A	141.38567A	157.09223A
3	2.730596	15.96642A	31.54522A	47.21009A	62.89651A	78.59154A	94.29088A	109.99269A	125.69604A	141.40041A	157.10550A
4	3.640326	16.14971A	31.63775A	47.27189A	62.94289A	78.62866A	94.32182A	110.01921A	125.71924A	141.42104A	157.12406A
5	4.549657	16.382436	31.75632A	47.35123A	63.00247A	78.67635A	94.36158A	110.05330A	125.74907A	141.44755A	157.14793A
6	5.458492	16.662640	31.90066A	47.44803A	63.07522A	78.73461A	94.41015A	110.09494A	125.78552A	141.47996A	157.17710A
7	6.366728	16.988042	32.07043A	47.56217A	63.16109A	78.80340A	94.46752A	110.14414A	125.82858A	141.51824A	157.21156A
8	7.274269	17.356215	32.26522A	47.69354A	63.26002A	78.88270A	94.53368A	110.20088A	125.87825A	141.56240A	157.25131A
9	8.181018	17.764617	32.484585	47.84200A	63.37197A	78.97248A	94.60860A	110.26515A	125.93452A	141.61244A	157.29636A
10	9.086883	18.210665	32.728057	48.00738A	63.49685A	79.07271A	94.69226A	110.33694A	125.99737A	141.66834A	157.34668A
11	9.991769	18.691814	32.995112	48.18953A	63.63460A	79.18333A	94.78465A	110.41623A	126.06681A	141.73010A	157.40229A
12	10.895584	19.205560	33.285178	48.38825A	63.78513A	79.30432A	94.88573A	110.50301A	126.14282A	141.79771A	157.46317A

TABLE 3
$\beta = 1.5$

s / n	1	2	3	4	5	6	7	8	9	10	11
1	0.805089	6.376507	12.61286A	18.88052A	25.15596A	31.43450A	37.71458A	43.99556A	50.27708A	56.55898A	62.84113A
2	1.608064	6.538074	12.69279A	18.93368A	25.19579A	31.46635A	37.74112A	44.01830A	50.29698A	56.57667A	62.85705A
3	2.406845	6.800078	12.82502A	19.02197A	25.26205A	31.51937A	37.78532A	44.05619A	50.33013A	56.60614A	62.88357A
4	3.199502	7.153187	13.008069	19.14495A	25.35455A	31.59347A	37.84711A	44.10917A	50.37651A	56.64737A	62.92069A
5	3.984302	7.586740	13.240072	19.30201A	25.47301A	31.68849A	37.92641A	44.17721A	50.43608A	56.70034A	62.96837A
6	4.759868	8.089989	13.518772	19.492366	25.61710A	31.80427A	38.02313A	44.26023A	50.50879A	56.76502A	63.02661A
7	5.525283	8.652747	13.841692	19.715153	25.78642A	31.94060A	38.13713A	44.35816A	50.59460A	56.84136A	63.09537A
8	6.280183	9.265874	14.206206	19.969378	25.980511	32.09722A	38.26828A	44.47091A	50.69344A	56.92934A	63.17463A
9	7.024764	9.920594	14.609666	20.253958	26.198870	32.27387A	38.41641A	44.59836A	50.80524A	57.02889A	63.26433A
10	7.759723	10.610011	15.049482	20.567770	26.440936	32.470234	38.58134A	44.74040A	50.92992A	57.13996A	63.36445A
11	8.486088	11.326711	15.523169	20.909628	26.706139	32.685987	38.76286A	44.89691A	51.06740A	57.26248A	63.47494A
12	9.205064	12.065020	16.028379	21.278336	26.993792	32.920781	38.96077A	45.06774A	51.21756A	57.39639A	63.59574A

TABLE 4
$\beta = 2.0$

s / n	1	2	3	4	5	6	7	8	9	10	11
1	0.677337	3.282470	6.35320A	9.47132A	12.60124A	15.73584A	18.87278A	22.01105A	25.15015A	28.28981A	31.42985A
2	1.340602	3.531292	6.474705	9.55158A	12.66121A	15.78374A	18.91265A	22.04521A	25.18003A	28.31636A	31.45374A
3	1.978877	3.920058	6.673802	9.684216	12.76068A	15.86331A	18.97896A	22.10205A	25.22976A	28.36057A	31.49353A
4	2.587614	4.418954	6.946140	9.867677	12.89893A	15.97417A	19.07148A	22.18142A	25.29925A	28.42236A	31.54916A
5	3.169444	4.992926	7.286813	10.100009	13.075002	16.11583A	19.18991A	22.28313A	25.38837A	28.50165A	31.62057A
6	3.731081	5.613492	7.690964	10.379013	13.287800	16.28765A	19.33386A	22.40693A	25.49695A	28.59832A	31.70767A
7	4.279317	6.254733	8.153854	10.702460	13.536067	16.488875	19.50289A	22.55254A	25.62478A	28.71221A	31.81035A
8	4.819109	6.898225	8.670590	11.068269	13.818503	16.718760	19.696490	22.71963A	25.77164A	28.84318A	31.92849A
9	5.353551	7.532315	9.233539	11.474624	14.133790	16.976439	19.914120	22.907796	25.93728A	28.99102A	32.06196A
10	5.884431	8.151579	9.834129	11.920106	14.480688	17.261049	20.155198	23.116669	26.12141A	29.15554A	32.21059A
11	6.412784	8.755433	10.460403	12.403537	14.858080	17.571711	20.419105	23.345816	26.32373A	29.33652A	32.37423A
12	6.939212	9.346029	11.100036	12.923681	15.265067	17.907606	20.705234	23.594788	26.54393A	29.53371A	32.55269A

TABLE 5

$\beta = 2.5$

s n	1	2	3	4	5	6	7	8	9	10	11
1	0.584712	2.263639	4.273300	6.33922A	8.41951A	10.50548A	12.59427A	14.68467A	16.77607A	18.86813A	20.96067A
2	1.136960	2.566404	4.422384	6.436661	8.49197A	10.56321A	12.64227A	14.72575A	16.81198A	18.90004A	20.98937A
3	1.643266	3.014111	4.666427	6.597438	8.611963	10.65901A	12.72201A	14.79406A	16.87173A	18.95314A	21.03715A
4	2.112840	3.541025	5.000385	6.819653	8.778449	10.792257	12.83313A	14.88935A	16.95515A	19.02731A	21.10392A
5	2.561246	4.090006	5.417247	7.101488	8.990216	10.962206	12.97514A	15.01131A	17.06200A	19.12239A	21.18956A
6	2.998819	4.626484	5.903285	7.441703	9.246103	11.168015	13.14745A	15.15952A	17.19202A	19.23818A	21.29391A
7	3.430580	5.139039	6.434476	7.839507	9.545276	11.40883A	13.34943A	15.33355A	17.34488A	19.37444A	21.41681A
8	3.858792	5.630601	6.981766	8.292655	9.887485	11.684016	13.580521	15.532939	17.52022A	19.53090A	21.55804A
9	4.284519	6.107923	7.522039	8.793645	10.273141	11.993085	13.840098	15.757181	17.717707	19.70728A	21.71739A
10	4.708335	6.576321	8.044625	9.327038	10.702781	12.336095	14.127730	16.005829	17.936898	19.903310	21.89462A
11	5.130594	7.038968	8.549352	9.872603	11.175293	12.713618	14.443191	16.278474	18.177451	20.11856A	22.08947A
12	5.551540	7.497588	9.040527	10.412958	11.684992	13.126721	14.786590	16.574836	18.439022	20.35285A	22.30169A

TABLE 6

$\beta = 3.0$

s n	1	2	3	4	5	6	7	8	9	10	11
1	0.513621	1.757764	3.236112	4.774941	6.32990A	7.89128A	9.45582A	11.02216A	12.58962A	14.15782A	15.72655A
2	0.977493	2.090108	3.406676	4.885065	6.411255	7.95586A	9.50941A	11.06797A	12.62963A	14.19335A	15.75850A
3	1.388031	2.542188	3.687198	5.067624	6.546166	8.063062	9.59844A	11.14413A	12.69619A	14.25247A	15.81167A
4	1.769223	3.019714	4.068117	5.322325	6.734053	8.212376	9.722532	11.25038A	12.78910A	14.33503A	15.88596A
5	2.137715	3.473757	4.520694	5.649721	6.974869	8.403356	9.88128A	11.38637A	12.90812A	14.44085A	15.98123A
6	2.500224	3.899285	4.998066	6.047230	7.269718	8.635923	10.074332	11.551802	13.052973	14.56971A	16.09731A
7	2.859234	4.306567	5.462847	6.499287	7.620584	8.910762	10.301571	11.746359	13.223370	14.721406	16.234020
8	3.215799	4.703812	5.903976	6.974669	8.026997	9.229644	10.563347	11.969928	13.419092	14.895667	16.391141
9	3.570477	5.095160	6.326594	7.443281	8.478551	9.595008	10.860782	12.222663	13.640002	15.092317	16.568487
10	3.923627	5.482493	6.737917	7.892390	8.951907	10.007154	11.195926	12.505215	13.886161	15.311226	16.76553A
11	4.275500	5.866719	7.142516	8.323863	9.421931	10.458332	11.571339	12.818914	14.157920	15.552398	16.98251A
12	4.626286	6.248353	7.542723	8.743594	9.875703	10.929875	11.987573	13.165952	14.456107	15.816020	17.21912A

TABLE 7
$\beta = 3.5$

s *n*	1	2	3	4	5	6	7	8	9	10	11
1	0.457116	1.454468	2.615215	3.837282	5.07677A	6.32324A	7.57314A	8.82498A	10.07803A	11.33188A	12.58630A
2	0.851944	1.796848	2.803828	3.957933	5.165169	6.393081	7.63093A	8.87430A	10.12106A	11.37006A	12.62061A
3	1.195733	2.218507	3.112756	4.160117	5.312582	6.509334	7.727080	8.95637A	10.19268A	11.43361A	12.67773A
4	1.518391	2.624121	3.510574	4.446164	5.520210	6.672168	7.861501	9.07104A	10.29275A	11.52241A	12.75757A
5	1.832864	2.996998	3.933706	4.811544	5.791144	6.882622	8.034366	9.218229	10.421139	11.63633A	12.86001A
6	2.143190	3.350464	4.336324	5.226722	6.128080	7.143320	8.246529	9.398148	10.577829	11.775335	12.984989
7	2.450808	3.694364	4.713594	5.645665	6.522097	7.458223	8.500044	9.611471	10.762977	11.939352	13.132395
8	2.756407	4.032861	5.075430	6.043274	6.942514	7.827645	8.798504	9.859776	10.977151	12.128544	13.302238
9	3.060412	4.367606	5.429045	6.420705	7.355345	8.237044	9.145286	10.145968	11.221559	12.343591	13.494652
10	3.363111	4.699379	5.777776	6.785951	7.748037	8.656638	9.536486	10.473877	11.498474	12.584795	13.710019
11	3.664716	5.028646	6.123040	7.144385	8.124571	9.062966	9.952988	10.844846	11.811287	12.854542	13.949078
12	3.965386	5.355740	6.465500	7.498594	8.491429	9.451361	10.368570	11.250050	12.163235	13.155423	14.213202

TABLE 8
$\beta = 4.0$

s *n*	1	2	3	4	5	6	7	8	9	10	11
1	0.411126	1.251118	2.202078	3.212742	4.241767	5.27817A	6.31824A	7.36038A	8.40380A	9.44806A	10.49292A
2	0.752324	1.589316	2.405800	3.342808	4.336187	5.352370	6.37944A	7.41249A	8.44920A	9.48830A	10.52905A
3	1.048387	1.966615	2.730106	3.563697	4.495331	5.476593	6.481572	7.499362	8.52486A	9.55534A	10.58925A
4	1.329098	2.310028	3.109408	3.875323	4.723462	5.652598	6.625298	7.621222	8.63082A	9.64917A	10.67349A
5	1.603863	2.627489	3.473731	4.247062	5.024741	5.884420	6.812303	7.778725	8.767350	9.769892	10.781799
6	1.875312	2.933211	3.810141	4.621730	5.385800	6.177375	7.046238	7.973473	8.935118	9.917781	10.914280
7	2.144462	3.232994	4.130526	4.971975	5.764014	6.527116	7.332444	8.208688	9.135594	10.093496	11.071230
8	2.411858	3.528861	4.443063	5.303097	6.124802	6.904143	7.671331	8.489368	9.371626	10.298392	11.253282
9	2.677860	3.821681	4.751048	5.624179	6.465090	7.272555	8.043947	8.818328	9.647736	10.534963	11.461681
10	2.942723	4.111963	5.055736	5.939872	6.793059	7.621048	8.417143	9.184515	9.967859	10.807262	11.698622
11	3.206628	4.400066	5.357709	6.252056	7.114497	7.955208	8.772954	9.559798	10.326486	11.119591	11.967695
12	3.469713	4.686272	5.657324	6.561496	7.432047	8.281445	9.113000	9.921876	10.701430	11.470218	12.273217

TABLE 9
$\beta = 4.5$

s n	1	2	3	4	5	6	7	8	9	10	11
1	0.373031	1.104180	1.907419	2.767026	3.645583	4.531899	5.42204A	6.31436A	7.20801A	8.10256A	8.99772A
2	0.672262	1.429128	2.122859	2.905834	3.745536	4.609929	5.486135	6.36881A	7.25537A	8.14448A	9.03533A
3	0.932730	1.761853	2.447073	3.143448	3.916412	4.741800	5.593741	6.459896	7.334463	8.21444A	9.09807A
4	1.181574	2.058531	2.788328	3.465085	4.164959	4.931864	5.746831	6.588549	7.44574A	8.31263A	9.18605A
5	1.425681	2.336965	3.099920	3.808756	4.484038	5.186972	5.949805	6.756880	7.590229	8.439631	9.299559
6	1.666949	2.607632	3.390988	4.128994	4.827895	5.503764	6.209304	6.969205	7.770095	8.596503	9.439164
7	1.906187	2.873843	3.672789	4.428087	5.153729	5.846841	6.524074	7.231840	7.989542	8.785455	9.606028
8	2.143874	3.136779	3.949702	4.716739	5.458655	6.176491	6.865922	7.544832	8.254520	9.010520	9.802340
9	2.380321	3.397053	4.223228	5.000127	5.752218	6.485955	7.198239	7.885246	8.565928	9.277316	10.031957
10	2.615753	3.655079	4.494004	5.280120	6.040102	6.783352	7.511375	8.219436	8.904830	9.587293	10.300197
11	2.850334	3.911170	4.762411	5.557443	6.324555	7.074637	7.811930	8.535612	9.240321	9.924668	10.608892
12	3.084192	4.165575	5.028733	5.832455	6.606423	7.362401	8.105941	8.838861	9.559061	10.261035	10.944743

TABLE 10
$\beta = 5.0$

s n	1	2	3	4	5	6	7	8	9	10	11
1	0.341023	0.992169	1.686615	2.433009	3.198648	3.972324	4.74998A	5.52992A	6.31124A	7.09349A	7.87637A
2	0.606945	1.298988	1.909895	2.579945	3.303850	4.053914	4.816710	5.586423	6.36030A	7.13684A	7.91523A
3	0.839813	1.592758	2.221209	2.829861	3.486480	4.193557	4.929637	5.681463	6.442497	7.209368	7.98018A
4	1.063471	1.854676	2.522036	3.142537	3.750957	4.398621	5.092836	5.817096	6.55894A	7.31164A	8.07151A
5	1.283122	2.103700	2.794283	3.445672	4.064092	4.672712	5.313819	5.997813	6.712042	7.445044	8.190073
6	1.500255	2.346950	3.053043	3.723205	4.367847	4.985908	5.594879	6.230930	6.906519	7.612121	8.337250
7	1.715570	2.586472	3.305774	3.987439	4.648429	5.289542	5.907917	6.517340	7.149364	7.817850	8.515920
8	1.929486	2.823103	3.554787	4.245629	4.916186	5.571978	6.211067	6.830058	7.440007	8.068763	8.731140
9	2.142291	3.057348	3.800915	4.500246	5.177982	5.842169	6.494648	7.132534	7.752285	8.362823	8.988886
10	2.354177	3.289571	4.044605	4.752137	5.436350	6.106535	6.766599	7.416815	8.053985	8.674564	9.285744
11	2.573055	3.520053	4.286170	5.001701	5.692160	6.367605	7.032887	7.690071	8.338672	8.975433	9.596870
12	2.775771	3.749018	4.525860	5.249210	5.945795	6.626274	7.296003	7.957848	8.612915	9.260326	9.896883

The roots calculated by the McMahon expression (3) are indicated in the tables by the letter A printed after the number. It should be noted that equation (3) does not give a root for $s = 1$.

The great speed of the IBM 704 digital computer was used to advantage in the solution of this problem. The Bessel functions were generated by using recursion relationships. Starting with an arbitrarily small number for a Bessel function of high order for a given argument, successively smaller orders were calculated until Bessel functions of 6- to 7-place accuracy were obtained. It was not possible with the particular procedure used to evaluate functions with arguments larger than 50.

The authors wish to express their appreciation for the advice given by Professor L. S. Han who originally suggested the problem.

Department of Mechanical Engineering
Ohio State University
Columbus, Ohio

1. J. McMahon, "On the roots of the Bessel and certain related functions," *Ann. of Math.*, v. 9, 1894, p. 23.

2. Rohn Truell, "Concerning the roots of $J_n'(x)N_n'(kx) - J_n'(kx)N_n'(x) = 0$," *J. Appl. Phys.*, v. 14, 1943, p. 350.

3. H. B. Dwight, "Table of roots for natural frequencies in coaxial type cavities," *J. Math. Phys.*, v. 27, 1948, p. 84.

Polynomial and Continued-Fraction Approximations for Logarithmic Functions

By Kurt Spielberg

1. Introduction. In this article we present the coefficients of approximations which are well suited for the calculation of logarithms on digital computers. The approximations have been derived by means of the IBM 704 program IB CTR. They are chosen so as to approximately minimize the absolute error over the appropriate interval of the argument. The method is described in detail in references [1], [2].

Similar selected polynomial approximations have been made available by C. Hastings [3]. The approximations of the present article, however, cover a much wider range of accuracy and should allow the coding of efficient double-precision subroutines.

Continued fraction approximations have been used systematically by E. G. Kogbetliantz and the author in connection with subroutines for the IBM 704 and 709 computers (see e.g. [4], [5], [6], which contain many references to other literature on rational approximations). The reader should note that the continued fraction approximations given in this paper not only allow for computation with fewer second-order arithmetic operations (multiplications and divisions) but also are intrinsically more accurate than polynomial approximations with equal numbers of constants.

2. Polynomial Approximations. In the case of digital computers, the argument can be assumed to be in normalized floating point form:

A. Binary machines:

$$y = 2^i \cdot f \tag{1.a}$$

$$i \cdots \text{integer}, \qquad f \cdots \text{fraction}, \qquad (\tfrac{1}{2}) \leqq f < 1.$$

B. Decimal machines:

$$y = 10^I \cdot F \tag{1.b}$$

$$I \cdots \text{integer}, \qquad F \cdots \text{fraction}, \qquad (\tfrac{1}{10}) \leqq F < 1.$$

The natural logarithm is then evaluated in accordance with the relations:

$$\log_e y = (i + \log_2 f) \cdot \log_e 2 \tag{2.a}$$

$$\log_e y = (I + \log_{10} F) \cdot \log_e 10. \tag{2.b}$$

To obtain efficient polynomial approximations, one starts with the well known series

$$\log_e \frac{v + x}{v - x} = 2[(x/v) + (x^3/3v^3) + (x^5/5v^5) + \cdots] \tag{3}$$

Received May 8, 1961

which converges in the interval $[-v < x < v]$. Since we intend to "economize" the power series by means of Chebyshev polynomials, we focus our attention on the interval $[-1 \leqq x \leqq 1]$. To this end we introduce the rational transformations

$$(4.a)\qquad f = u \cdot \left(\frac{v + x}{v - x}\right), \qquad x = v \cdot \left(\frac{f - u}{f + u}\right)$$

$$(4.b)\qquad F = U \cdot \left(\frac{V + x}{V - x}\right), \qquad x = V \cdot \left(\frac{F - U}{F + U}\right)$$

and determine u, v, U, V so that the interval $[-1 \leqq x \leqq 1]$ maps one-to-one onto the intervals $[(\frac{1}{2}) \leqq f \leqq 1]$ and $[(\frac{1}{10}) \leqq F \leqq 1]$ respectively.

The parameters are determined from the endpoint conditions:

$$(5.a)\qquad u \cdot \left(\frac{v + 1}{v - 1}\right) = 1, u \cdot \left(\frac{v - 1}{v + 1}\right) = \frac{1}{2} \quad \rightarrow u = 1/\sqrt{2},$$
$$v = (\sqrt{2} + 1)^2$$

$$(5.b)\qquad U \cdot \left(\frac{V + 1}{V - 1}\right) = 1, U \cdot \left(\frac{V - 1}{V + 1}\right) = \frac{1}{10} \quad \rightarrow U = 1/\sqrt{10},$$
$$V = (\sqrt{10} + 1)^2/9.$$

On substituting these values into equation 3, we obtain the following power series for $\log_2 f$ and $\log_{10} F$:

$$(6.a)\qquad \log_2 f = 2 \cdot \log_2 e \cdot [(x/v) + (x^3/3v^3) + \cdots] - (\tfrac{1}{2})$$

$$(6.b)\qquad \log_{10} F = 2 \cdot \log_{10} e \cdot [(x/V) + (x^3/3V^3) + \cdots] - (\tfrac{1}{2}), \qquad [-1 \leqq x \leqq 1]$$

The 704 program IB CTR is now applied to produce polynomial approximations to the functions $\log_2 f + (\frac{1}{2})$ and $\log_{10} F + (\frac{1}{2})$. These approximations have the form:

$$f_m^*(x) = \sum_{i=1}^{m} \bar{c}_{2i-1}^{(m)} \cdot x^{2i-1}$$

$$(7)\qquad x = v \frac{f - (\sqrt{2}/2)}{f + (\sqrt{2}/2)} \quad \text{for} \quad \log_2 f + \left(\frac{1}{2}\right)$$

$$x = V \frac{F - (\sqrt{10}/10)}{F + (\sqrt{10}/10)} \quad \text{for} \quad \log_{10} F + \left(\frac{1}{2}\right).$$

For computational purposes, however, it is preferable to introduce the variables $z = \dfrac{x}{v}$ or $z = \dfrac{x}{V}$,

$$(8)\qquad f_m^*(z) = \sum_{i=1}^{m} c_{2i-1}^{(m)} \cdot z^{2i-1}.$$

In Tables 1 and 2 we give the coefficients $c_j^{(m)}$ for those m which result in approximations of less than or equal to 16-digit accuracy. IB CTR performs operations to 16-digit accuracy only. Its primary output, however, consists of the increments Δa_i which have to be added to the power series coefficients a_i of the given function to produce the coefficients of the approximation polynomial. We therefore give

Table 1

Polynomial Coefficients for $(Log_2 f + \frac{1}{2})$

[Format: 451 E-16 = (.451) × 10^{-16}]

# of coefficients = 2, E1 = 606 E-05,	E2 = 560 E-05, E3 = 876 E-04			
2885 2290 8878 0725 E01	9835 1099 0073 3683 E00			
# of coefficients = 3, E1 = 317 E-07,	E2 = 299 E-07, E3 = 185 E-05			
2885 3912 8434 1961 E01	9614 7149 2111 3942 E00	5989 5531 8769 7430 E00		
# of coefficients = 4, E1 = 182 E-09,	E2 = 174 E-09, E3 = 423 E-07			
2885 3900 7279 5173 E01	9618 0074 7520 7173 E00	5765 8533 5684 3871 E00	4342 4052 2333 9827 E00	
# of coefficients = 5, E1 = 111 E-11,	E2 = 106 E-11, E3 = 102 E-08			
2885 3900 8184 5024 E01	9617 9664 8473 7566 E00	5770 8662 4639 5350 E00	4115 3509 8458 0017 E00	3428 0712 2893 2386 E00
# of coefficients = 6, E1 = 699 E-14,	E2 = 680 E-14, E3 = 254 E-10			
2885 3900 8177 7425 E01	9617 9669 4401 5285 E00	5770 7788 7562 8954 E00	4122 1350 6252 1815 E00	3197 6122 8252 9226 E00
2846 8436 5237 6334 E00				
# of coefficients = 7, E1 = 451 E-16,	E2 = 111 E-15, E3 = 649 E-12			
2885 3900 8177 7930 E01	9617 9669 3921 2376 E00	5770 7801 8088 9661 E00	4121 9830 3814 7730 E00	3206 2195 4010 7357 E00
2612 9289 6508 8097 E00	2444 9890 0547 1920 E00			

Table 1′

Increments to Polynomial Coefficients for $(Log_2 f + \frac{1}{2})$

# of coefficients = 8, E1 = 296 E-18,	E3 = 168647 E-13			
−2796 4430 3552 4398 E-16	4551 4828 0666 6136 E-13	−2159 4856 0182 7467 E-10	4596 5458 7348 9827 E-08	−5181 9115 5104 8872 E-06
3306 6196 2511 6075 E-04	−1196 4114 9893 8173 E-02	2271 7162 0774 9647 E-01		
# of coefficients = 9, E1 = 197 E-20,	E3 = 444326 E-15			
2088 8319 0383 6615 E-18	−4250 6312 1996 8916 E-15	2536 1559 2048 5863 E-12	−6874 1931 2702 1155 E-10	1006 3607 9885 3810 E-07
−8676 6985 8599 0271 E-06	4501 4264 0509 3858 E-04	−1381 0846 5746 4859 E-02	2289 0231 2019 8052 E-01	
# of coefficients = 10, E1 = 133 E-22,	E3 = 118370 E-16			
−1560 2706 2751 6011 E-20	3881 2970 6574 1623 E-17	−2842 8056 6023 4796 E-14	9543 6054 6158 3396 E-12	−1759 9844 9576 6640 E-09
1949 1463 1860 0759 E-07	−1350 9195 6439 6997 E-05	5896 2316 8612 9454 E-04	−1560 3718 5748 8830 E-02	2306 5044 3068 2277 E-01
# of coefficients = 11, E1 = 906 E-25,	E3 = 318215 E-18			
1165 4539 6029 6268 E-22	−3479 5084 3296 5673 E-19	3068 2604 2428 8200 E-16	−1248 3015 3206 5234 E-13	2820 0813 2649 2660 E-11
−3889 3504 0329 2187 E-09	3441 0813 3886 3375 E-07	−1992 1520 3031 4775 E-05	7497 3891 8807 8451 E-04	−1761 3336 7553 1071 E-02
2324 1621 0740 7857 E-01				
# of coefficients = 12, E1 = 622 E-27,	E3 = 861954 E-20			
−8706 4029 5067 7987 E-25	3071 9811 7588 6013 E-21	−3209 5080 2494 8594 E-18	1554 8328 4036 6698 E-15	−4216 4739 5337 7175 E-13
7064 8242 3847 3655 E-11	−7727 2510 5334 2210 E-09	5674 1027 1296 3337 E-07	−2816 6347 8303 9504 E-05	9311 4020 8001 2885 E-04
−1957 0816 5775 8419 E-02	2341 9981 4509 0346 E-01			

TABLE 2
Polynomial Coefficients for $(Log_{10} F + \frac{1}{2})$

# of coefficients = 2, E1 = 687 E-03,	E2 = 635 E-03, E3 = 818 E-02			
8632 4020 7770 2620 E00	3631 7894 3815 3434 E00			
# of coefficients = 3, E1 = 380 E-04,	E2 = 358 E-04, E3 = 161 E-02			
8690 0869 5758 3972 E00	2776 7991 9776 1202 E00	2534 4884 5752 8303 E00		
# of coefficients = 4, E1 = 230 E-05,	E2 = 219 E-05, E3 = 342 E-03			
8685 5602 6257 9886 E00	2910 9861 8255 4689 E00	1540 0462 1701 3371 E00	2105 6233 5032 9898 E00	
# of coefficients = 5, E1 = 147 E-06,	E2 = 141 E-06, E3 = 763 E-04			
8685 9154 8508 7103 E00	2893 4361 3145 7315 E00	1774 1562 6792 7102 E00	9489 8398 4169 7810 E-01	1904 8229 9261 9342 E00
# of coefficients = 6, E1 = 971 E-08,	E2 = 938 E-08, E3 = 175 E-04			
8685 8876 0977 6415 E00	2895 5019 3437 1026 E00	1731 2899 5934 1548 E00	1312 0429 7994 7967 E00	5595 3156 3065 2789 E-01
1812 6871 4549 2145 E00				
# of coefficients = 7, E1 = 659 E-09,	E2 = 639 E-09, E3 = 412 E-05			
8685 8897 9723 0560 E00	2895 2749 8124 8346 E00	1738 0176 3929 2238 E00	1226 5721 9388 9737 E00	1087 3756 0172 2274 E00
2479 7381 8779 1450 E-01	1783 9825 1660 6317 E00			
# of coefficients = 8, E1 = 455 E-10,	E2 = 443 E-10, E3 = 986 E-06			
8685 8896 2557 4872 E00	2895 2987 2744 9401 E00	1737 0673 4730 0434 E00	1243 3400 2720 4158 E00	9354 9710 3959 7025 E-01
9846 9859 1034 3788 E-01	−3594 1212 1911 1870 E-02	1798 2956 4764 1127 E00		
# of coefficients = 9, E1 = 319 E-11,	E2 = 311 E-11, E3 = 239 E-06			
8685 8896 3904 5177 E00	2895 2963 3160 9180 E00	1737 1916 3411 5875 E00	1240 4452 2516 2623 E00	9712 5210 4496 9738 E-01
7341 7616 3051 0283 E-01	9637 6057 5788 3042 E-01	−3184 6422 2742 1276 E-01	1845 5330 9203 7982 E00	
# of coefficients = 10, E1 = 226 E-12,	E2 = 221 E-12, E3 = 585 E-07			
8685 8896 3798 8125 E00	2895 2965 6661 9593 E00	1737 1763 0775 8544 E00	1240 8995 5170 2479 E00	9639 5703 5930 4128 E-01
8029 8336 5029 0451 E-01	5715 1350 0078 7728 E-01	1010 4006 5135 6441 E00	−6164 7796 8076 9668 E-01	1920 5939 9039 2521 E00
# of coefficients = 11, E1 = 162 E-13,	E2 = 159 E-13, E3 = 144 E-07			
8685 8896 3807 1075 E00	2895 2965 4407 9458 E00	1737 1781 1181 2187 E00	1240 8333 4031 9255 E00	9652 9278 6613 0788 E-01
7867 8488 0187 9711 E-01	6946 3643 1539 0860 E-01	4194 8195 6636 4441 E-01	1122 3305 1516 4517 E00	−9429 7998 7036 6938 E-01
2021 1022 0368 1399 E00				
# of coefficients = 12, E1 = 117 E-14,	E2 = 117 E-14, E3 = 359 E-08			
8685 8896 3806 4564 E00	2895 2965 4620 2003 E00	1737 1779 0732 3385 E00	1240 8424 3295 8296 E00	9650 6817 2073 5434 E-01
7901 7458 8805 0730 E-01	6617 8630 0558 8470 E-01	6281 5170 1450 0886 E-01	2581 5225 2642 8773 E-01	1304 1471 6289 5274 E00
−1309 5310 3342 3318 E00	2146 3381 2708 5184 E00			

# of coefficients = 13, E1 = 847 E-16,	E2 = 111 E-15,	E3 = 900 E-09		
8685 8896 3806 5076 E00	2895 2965 4600 5157 E00	1737 1779 2978 8923 E00	1240 8412 4374 9968 E00	9651 0342 4344 7121 E-01
7895 2858 8846 5882 E-01	6695 1982 7477 7399 E-01	5659 2707 1451 3706 E-01	5972 2476 4507 6175 E-01	6077 8384 6561 5379 E-02
1565 4748 5242 4707 E00	−1727 5334 3561 2921 E00	2296 7005 4645 3325 E00		
# of coefficients = 14, E1 = 662 E-17,	E2 = 555 E-16,	E3 = 226 E-09		
8685 8896 3806 5035 E00	2895 2965 4602 3178 E00	1737 1779 2738 4921 E00	1240 8413 9306 2811 E00	9650 9819 9026 3539 E-01
7896 4264 9078 3005 E-01	6678 7260 4454 3685 E-01	5822 0354 4727 4206 E-01	4854 7126 9236 7506 E-01	6025 3255 6326 7152 E-01
−1642 0035 6283 8796 E-01	1920 3899 2012 4037 E00	−2209 0134 7726 2696 E00	2473 4265 8441 2308 E00	

TABLE 2′

Increments to Polynomial Coefficients for ($Log_{10} F + ½$)

```
# of coefficients = 13,   E1 = 847  E-16,   E2 = 111  E-15,   E3 = 899533  E-09
     4008 4239 3911 5484 E-14    -1803 1583 5441 1154 E-11     2404 0008 8731 8954 E-09    -1493 1284 3346 2257 E-07     5225 3183 5780 7491 E-06
    -1140 6023 1713 0290 E-04     1647 2230 2337 1189 E-03    -1627 6473 2760 5324 E-02     1117 5349 5270 8717 E-01    -5327 5417 1670 5883 E-01
     1729 6752 0870 8676 E00     -3647 9233 5573 7152 E00      4505 7140 2371 7118 E00

# of coefficients = 14,   E1 = 662  E-17,   E2 = 555  E-16,   E3 = 226368  E-09
    -2916 4626 3720 0172 E-15     1497 3893 0252 3668 E-12    -2280 8840 3338 6883 E-10     1619 6375 3240 7678 E-08    -6496 4754 0810 2044 E-07
     1631 8345 1136 6819 E-05    -2727 5231 9866 8126 E-04     3144 2355 2308 3100 E-03    -2546 3415 3553 1193 E-02     1453 8047 0112 7640 E-01
    -5778 1414 8572 6928 E-01     1542 7425 4455 5990 E00     -2556 4490 6278 5299 E00      2151 7269 6818 7675 E00

# of coefficients = 15,   E1 = 486  E-18,   E3 = 572275  E-10
     2288 5535 9684 9383 E-16    -1343 8391 4610 9992 E-13     2344 3219 5244 5695 E-11    -1911 7146 2297 6515 E-09     8843 4268 4050 7464 E-08
    -2577 4243 7703 7674 E-06     5040 3255 9895 0395 E-05    -6875 4658 7637 3031 E-04     6691 8269 8604 0228 E-03    -4693 0443 9505 2550 E-02
     2365 8076 4831 7931 E-01    -8427 8096 1729 4303 E-01     2040 0398 5490 6148 E00     -3088 8853 0687 4070 E00      2378 9844 2405 9968 E00

# of coefficients = 16,   E1 = 358  E-19   E3 = 145258  E-10
    -1795 9736 0410 3296 E-17     1195 8805 9185 7760 E-14    -2368 5279 2028 3146 E-12     2197 7461 1744 1895 E-10    -1160 8259 9300 6863 E-08
     3881 8413 3500 7086 E-07    -8767 5766 2421 1050 E-06     1393 5293 3967 7936 E-04    -1599 0567 3726 0520 E-03     1343 4272 4682 2011 E-02
    -8295 8548 8882 7127 E-02     3740 1090 5421 8566 E-01    -1206 2550 3494 5501 E00      2668 5531 8199 7598 E00     -3716 4406 6478 1781 E00
     2634 1345 0733 5488 E00

# of coefficients = 17,  E1 = 265  E-20,   E3 = 370012  E-11
     1409 3150 8547 3774 E-18    -1056 3115 8573 2524 E-15     2357 1276 8150 9970 E-13    -2468 8104 8591 3208 E-11     1476 1050 4700 8316 E-09
    -5609 8840 3888 0974 E-08     1447 0875 5677 5826 E-06    -2647 4854 3802 0277 E-05     3527 7025 8319 1047 E-04    -3483 5465 5112 1759 E-03
     2570 0434 3901 5269 E-02    -1416 0855 4267 2644 E-01     5769 6629 1498 4858 E-01    -1698 8701 6376 2035 E00      3458 3781 1427 7836 E00
    -4455 4183 0129 9020 E00      2921 0185 6128 9769 E00

# of coefficients = 18,   E1 = 196  E-21,   E3 = 945511  E-12
    -1105 9688 8198 7785 E-19     9268 7827 8567 9632 E-17    -2314 5895 1668 1069 E-14     2717 1493 4957 1405 E-12    -1825 1852 6724 5195 E-10
     7818 7429 4464 7917 E-09    -2284 2814 5794 9697 E-07     4756 2823 4455 0482 E-06    -7268 9080 6071 2682 E-05     8311 1740 8052 1467 E-04
    -7188 5605 8617 9683 E-03     4721 4631 9690 9278 E-02    -2346 5322 9941 6829 E-01     8715 5992 3026 6765 E-01    -2359 5803 5285 6509 E00
     4445 9765 4492 1987 E00     -5324 8659 4849 2503 E00      3243 7845 5215 9573 E00

# of coefficients = 19,   E1 = 146  E-22,   E3 = 242297  E-12
     8678 6815 8826 4611 E-21    -8085 0793 9863 8296 E-18     2245 8153 1814 8675 E-15    -2936 4513 5611 5064 E-13     2201 3687 5694 3754 E-11
    -1055 3520 0829 8306 E-09     3463 0894 7261 7532 E-08    -8137 0689 8363 0670 E-07     1411 7100 4334 4387 E-05    -1846 3409 3646 6263 E-04
     1844 5616 7118 6270 E-03    -1417 2101 0408 5692 E-02     8379 7119 6535 3521 E-02    -3790 3518 8690 6554 E-01     1292 8323 5304 1614 E00
    -3237 7162 4057 8437 E00      5675 4563 7087 2527 E00     -6347 0335 4087 0700 E00      3607 2084 8706 5067 E00

# of coefficients = 20,   E1 = 109  E-23,   E3 = 622496  E-13
    -6810 2228 2251 4817 E-22     7015 1294 5830 8251 E-19    -2155 8115 2234 0397 E-16     3121 9959 4939 8246 E-14    -2596 6481 6481 3679 E-12
     1384 3409 4148 1999 E-10    -5067 1511 6295 6044 E-09     1333 2738 1946 9607 E-07    -2603 0804 1617 8951 E-06     3855 0774 2268 1785 E-05
    -4395 2520 2266 6984 E-04        3892 3608 8149 6460 E-03       -2687 1533 6395 1063 E-02        1443 6598 9658 1256 E-01       -5988 2739 0797 6818 E-01
     1887 4523 9004 8467 E00        -4395 5168 5808 9775 E00         7200 0996 3927 2206 E00        -7547 9067 2022 7777 E00         4016 7260 9603 3941 E00

# of coefficients = 21,    E1 = 816  E-25,    E3 = 160299  E-13

     5344 3892 5310 2525 E-23       -6057 5655 6179 9451 E-20        2049 4387 4450 8078 E-17       -3270 6768 7331 9013 E-15        3002 1538 1713 7558 E-13
    -1760 8835 7932 5375 E-11        7182 5272 3571 8276 E-10       -2102 2162 7827 6335 E-08        4584 3885 2468 5963 E-07       -7622 3540 8301 8407 E-06
     9819 0989 1950 0592 E-05       -9903 7750 4168 7679 E-04        7866 3901 2604 2533 E-03       -4926 2758 2856 7780 E-02        2423 5452 9496 9185 E-01
    -9278 5253 8561 7140 E-01        2717 1618 6520 7351 E00        -5911 2413 0888 5684 E00         9084 1861 8185 0081 E00        -8957 8287 4109 7084 E00
     4478 5262 7490 0550 E00

# of coefficients = 22,    E1 = 612  E-26,    E3 = 413655  E-14

    -4193 8089 2404 0763 E-24        5207 9021 4801 2185 E-21       -1931 2330 2346 2839 E-18        3380 9297 1741 7593 E-16       -3408 6150 5610 5388 E-14
     2210 9742 3128 4624 E-12       -9894 1765 6093 8458 E-11        3202 3242 6442 6305 E-09       -7749 4621 9111 2255 E-08        1435 9844 9702 4511 E-06
    -2072 5304 1843 1378 E-05        2357 4164 4914 3829 E-04       -2128 8798 5630 2663 E-03        1531 4262 9655 0943 E-02       -8760 5043 5061 7090 E-02
     3976 9690 6299 1939 E-01       -1413 1624 4153 8475 E00         3863 1553 0659 5049 E00        -7882 9871 1609 8231 E00         1140 5168 5158 9211 E01
    -1061 2225 3706 5133 E01         4999 6580 0646 9901 E00

# of coefficients = 23,    E1 = 457  E-27,    E3 = 106950  E-14

     3291 1012 0129 4949 E-25       -4459 5968 1784 8354 E-22        1805 2889 8664 4529 E-19       -3452 5888 7283 9814 E-17        3806 8023 8461 8771 E-15
    -2704 5103 1627 6891 E-13        1328 1452 5263 4629 E-11       -4728 7227 5666 3148 E-10        1262 5826 3924 1371 E-08       -2590 7962 4956 1663 E-07
     4150 2197 4371 8105 E-06       -5290 9424 1824 6419 E-05        5379 3390 1236 8034 E-04       -4392 8399 1546 7543 E-03        2885 4594 6430 3172 E-02
    -1521 2097 0908 0803 E-01        0395 8604 1552 0965 E-01       -2119 0352 6547 6698 E00         5431 6316 5652 5391 E00        -1043 3368 7038 7462 E01
     1425 6264 1187 6378 E01        -1255 2449 7724 7013 E01         5588 1527 2783 5117 E00

# of coefficients = 24,    E1 = 320  E-28,    E3 = 277007  E-15

    -2582 5814 8673 8534 E-26        3804 8849 0235 4618 E-23       -1678 1940 2563 1511 E-20        3486 6979 1696 7982 E-18       -4187 9243 2519 6063 E-16
     3245 3634 1605 9773 E-14       -1741 3525 5568 6211 E-12        0788 3390 1234 3116 E-11       -1989 0513 4147 8110 E-09        4405 7966 2581 7089 E-08
    -7977 8302 5787 6452 E-07        1126 9152 1233 6463 E-05       -1279 3277 9037 5553 E-04        1174 4587 7238 1308 E-03       -8745 5383 9347 7507 E-03
     5282 1375 3009 4918 E-02       -2578 8727 3323 3044 E-01        1010 2742 2496 2924 E00        -3136 0304 9913 4927 E00         7560 8121 7592 3551 E00
    -1371 5241 6034 4965 E01         1774 9543 3483 0017 E01        -1482 6767 1200 5955 E01         6253 1643 9083 9764 E00
```

tables of these increments from which the reader can construct approximations of great accuracy by simple hand computation.

All approximations of the form (7) have been tested at more than 100 points in the interval $[-1 \leqq x \leqq 1]$. Instead of the complete error curves we submit, for simplicity, three "error parameters."

$E_1 \cdots$ a theoretical upper bound of the magnitude of the absolute error caused by a truncation of a Chebyshev series to m terms

$E_2 \cdots$ the maximum magnitude of the absolute error encountered in the described test

$E_3 \cdots \sum_{i=m+1}^{\infty} a_{2i-1}$, the maximum absolute error incurred by a truncation of the given power series to m terms.

The sets of increments have been tested as follows. From the definitions we infer that (for $x = 1$)

$$\sum_{i=1}^{m} a_{2i-1} + \sum_{i=m+1}^{\infty} a_{2i-1} = \sum_{i=1}^{m} (a_{2i-1} + \Delta a_{2i-1}) \pm \max (E_1, E_2)$$

or

$$E_3 = \sum_{i=1}^{m} \Delta a_{2i-1} \pm \max (E_1, E_2).$$

Selected tests of this type have consistently been satisfactory. The reader should note, however, that these tests do not usually apply to the last two digits due to the unfortunate fact that E_3 has been printed only to 6 digits. In order to obtain a better check, at least up to "triple precision accuracy" on the IBM 704 (2^{-70}), we have therefore coded a triple precision logarithm subroutine based on the given increments. The accuracy of the subroutine was verified by an application to functional relationships of the form $\log (x \cdot y) = \log x + \log y$. We have every reason to believe that all of the given increments will be found to be completely accurate.

3. Continued Fraction Approximations. An approximation polynomial can be transformed into a rational approximation with the same number of constants by means of the "multiple truncation procedure" described in [2] and implemented in IB CTR. It is shown in [2] that the rational approximation may actually be considerably better than the original polynomial approximation. The results submitted in the present article furnish an excellent instance of this behavior.

Rational approximations can readily be transformed into continued fractions which can be evaluated in fewer operations. In Tables 3 and 4 we give the continued fraction expressions for $(\log_2 f + \frac{1}{2})$ and $(\log_{10} F + \frac{1}{2})$ up to 16-digit accuracy. They are of the form

$$g_m^*(z)/z = H_0 + \frac{G_1 \;|}{|\; z^2 + H_1} + \frac{G_2 \;|}{|\; z^2 + H_2} + \cdots + \frac{G_{[m/2]} \;|}{|\; z^2 + H_{[m/2]}},$$

where $m = 3, 4, \cdots$ and $[\frac{1}{2}m]$ is the largest integer $\leqq \frac{m}{2}$. For even m, the constant H_0 is zero.

TABLE 3

Continued Fraction Coefficients for $(Log_2\ F + \frac{1}{2})$

# of coefficients = 3, E1 = 482 E-08,	E2 = 479 E-08, E3 = 185 E-05		
1292 0070 9870 0440 E01	−1656 7626 3013 4752 E01		
−2639 8577 0311 1530 E01			
# of coefficients = 4, E1 = 719 E-11,	E2 = 971 E-11, E3 = 423 E-07		
0000 0000 0000 0000 00	−7987 3541 1394 9024 E01	−1893 0810 4263 2588 E01	
−1747 9113 9907 0586 E02	−3652 8036 5468 0985 E01		
# of coefficients = 5, E1 = 577 E-14,	E2 = 226 E-13, E3 = 102 E-08		
8270 7235 6521 4108 E00	−3085 7167 7071 2198 E01	−1540 1793 1703 5943 E01	
−5536 8085 7347 3842 E01	−6095 2425 1953 1276 E00		
# of coefficients = 6, E1 = 794 E-16,	E2 = 500 E-15, E3 = 254 E-10		
0000 0000 0000 0000 00	−1561 2646 7876 6829 E02	−3704 0518 2875 6762 E01	−1390 3458 4580 7226 E01
−2641 9662 6270 7515 E02	−2281 0512 9374 2221 E02	−2377 2595 4404 4343 E00	
# of coefficients = 7, E1 = 100 E-17,	E2 = 167 E-15, E3 = 649 E-12		
6089 0779 0129 1574 E00	−5026 9809 0578 5215 E01	−2550 5195 2771 6355 E01	−1306 2235 3597 6247 E01
−8465 8394 7460 5167 E01	−3212 8370 0542 7867 E01	−1234 5730 6822 9921 E00	

Constants are listed in the following sequence:
First line (or lines): $H_0, H_1, \cdots, H_{[m/2]}$
New line: $G_1, G_2, \cdots, G_{[m/2]}$

TABLE 4

Continued Fraction Coefficients for (Log₁₀ F + ½)

# of coefficients = 3, E1 = 271 E-05,	E2 = 587 E-05, E3 = 161 E-02			
4174 8684 7078 8742 E00	−1567 8846 4453 5199 E01			
−7073 8930 5767 5742 E00				
# of coefficients = 4, E1 = 314 E-06,	E2 = 184 E-06, E3 = 342 E-03			
0000 0000 0000 0000 00	−7006 1359 6128 1824 E01	−1741 4420 1618 7053 E01		
−4806 7940 1892 9581 E01	−2563 5634 1028 4778 E01			
# of coefficients = 5, E1 = 434 E-07,	E2 = 877 E-08, E3 = 763 E-04			
2735 8316 6302 7741 E00	−2733 5680 1437 2866 E01	−1431 6382 7957 1771 E01		
−1460 7439 0859 0961 E01	−3987 9825 1331 3200 E00			
# of coefficients = 6, E1 = 497 E-08,	E2 = 744 E-09, E3 = 175 E-04			
0000 0000 0000 0000 00	−1217 4701 9298 6194 E02	−2958 5851 1583 7019 E01	−1269 2331 5593 5871 E01	
−6828 6213 0481 5603 E01	−1234 6312 3127 7424 E02	−1218 1864 1232 5462 E00		
# of coefficients = 7, E1 = 605 E-09,	E2 = 589 E-10, E3 = 412 E-05			
2257 0476 5469 1600 E00	−3630 0459 1190 2254 E01	−1860 6441 8950 3682 E01	−1081 7612 3251 9165 E01	
−1928 2949 9018 7036 E01	−1158 3549 7192 2684 E01	−2568 2365 4014 4450 E-01		
# of coefficients = 8, E1 = 119 E-09,	E2 = 481 E-11, E3 = 986 E-06			
0000 0000 0000 0000 00	−1552 9064 0908 9411 E02	−3714 7223 4968 7198 E01	−1412 1175 9204 4839 E01	−7168 7986 8486 6398 E00
−7920 3098 1162 8075 E01	−2267 8348 7376 9289 E02	−2498 4529 6288 2307 E00	−5063 2900 8671 8532 E-03	
# of coefficients = 9, E1 = 156 E-10,	E2 = 623 E-12, E3 = 239 E-06			
1925 4014 9937 8824 E00	−4664 6742 0574 2049 E01	−2390 9878 0081 [illegible] E01	−1274 9458 0829 0809 E01	−3522 2956 7225 2972 E00
−2393 3428 1057 6800 E01	−2600 4549 6660 8233 E01	−9996 4987 5811 9110 E-01	−3358 7524 7535 1047 E-07	
# of coefficients = 10, E1 = 434 E-12,	E2 = 156 E-11, E3 = 585 E-07			
0000 0000 0000 0000 00	−2278 8796 3815 2385 E02	−5544 3240 6207 0304 E01	−1934 3441 8577 6231 E01	−1218 2064 2048 6353 E01
−1897 0504 6669 6781 E00	0000 0000 0000 0000 00	0000 0000 0000 0000 00	0000 0000 0000 0000 00	0000 0000 0000 0000 00
−9900 4585 3391 4664 E01	−5795 9838 5747 5431 E02	−8607 6714 8406 6327 E00	−5625 5676 2020 1057 E-01	−7055 7921 1811 2189 E-12
# of coefficients = 11, E1 = 254 E-13,	E2 = 603 E-13, E3 = 144 E-07			
1566 4660 0628 8101 E00	−6592 2236 1492 1441 E01	−3455 8641 8980 6945 E01	−1697 0477 1482 5547 E01	−1180 8990 8952 5397 E01
−1369 5502 7639 1718 E00	0000 0000 0000 0000 00	0000 0000 0000 0000 00	0000 0000 0000 0000 00	0000 0000 0000 0000 00
−3126 5747 7180 0882 E01	−7074 7971 1034 8246 E01	−4016 2950 6095 0182 E00	−3557 5429 4422 7386 E-01	−1782 4702 3079 4942 E-12
# of coefficients = 12, E1 = 187 E-14,	E2 = 334 E-13, E3 = 359 E-08			
0000 0000 0000 0000 00	−3169 9223 8947 9199 E02	−7831 1526 9137 3795 E01	−2568 4206 7268 4757 E01	−1525 7373 3977 9744 E01
−1144 0138 7296 9406 E01	−1201 4681 1850 7815 E00	0000 0000 0000 0000 00	0000 0000 0000 0000 00	0000 0000 0000 0000 00

```
    −1193 3840 8458 0869 E02     −1251 0738 8678 6443 E03     −2104 3075 1900 4511 E01     −2053 9093 1711 4612 E00     −2160 1176 0348 0808 E-01
    −5688 4767 1640 4652 E-12

# of coefficients = 13,    E1 = 174  E-15,    E2 = 149  E-13,    E3 = 900  E-09
     1393 1454 3501 9793 E00     −8062 3888 3507 8176 E01     −4276 3916 4675 1179 E01     −2007 5909 6834 5058 E01     −1343 0501 1722 4303 E01
    −1061 3732 5582 5836 E01     −1130 8536 1647 6820 E00      0000 0000 0000 0000  00      0000 0000 0000 0000  00      0000 0000 0000 0000  00
    −3615 1063 3147 0115 E01     −1206 7169 2646 9193 E02     −7573 1788 7233 3270 E00     −9019 7116 9776 3227 E-01    −7211 9042 7185 6326 E-02
    −2655 8838 7827 0656 E-11

# of coefficients = 14,    E1 = 867  E-17,    E2 = 153  E-13,    E3 = 226  E-09
     0000 0000 0000 0000  00     −3693 2105 0974 0172 E02     −9153 8599 6300 3995 E01     −2900 6086 3002 2923 E01     −1656 2571 8132 1678 E01
    −1191 9233 8543 8025 E01     −7551 8137 2851 1484 E00     −1068 6185 0036 1872 E00      0000 0000 0000 0000  00      0000 0000 0000 0000  00
    −1299 9443 2325 4275 E02     −1766 0355 0332 9192 E03     −3020 2074 0796 5379 E01     −3080 7913 8786 6877 E00     −3556 1104 2194 9375 E-01
    −6747 9920 9079 2036 E-04    −2239 5413 9687 9253 E-09
```

All continued fractions have been checked at more than 100 points in the interval $[-1 \leqq x \leqq 1]$. These checks were executed in double precision arithmetic, i.e., with wordlengths of 16 digits. The user of a particular continued fraction approximation should briefly analyze how much round-off error may accrue on his machine due to limited wordlength and subtraction of numbers of equal magnitude. For large H_i and G_i serious loss of accuracy might occur in this manner. In the case of the logarithmic functions, however, little difficulty should arise from this source.

4. Use of Tables. To illustrate the use of the tables we give a few simple examples.

a) Polynomial approximation, three coefficients:

$$\frac{1}{2} \leqq f \leqq 1, \qquad z = \frac{f - \frac{\sqrt{2}}{2}}{f + \frac{\sqrt{2}}{2}}$$

$$\log_2 f = f_3^*(z) - \tfrac{1}{2} \pm (.32)\ .10^{-7} = c_1 z + c_3 z^3 + c_5 z^5 - \tfrac{1}{2} \pm (.32)\ .10^{-7}$$

Table 1:

$$c_1 = 2.88539 \quad 12843 \cdots$$
$$c_3 = .96147 \quad 14921 \cdots$$
$$c_5 = .59895 \quad 53187 \cdots$$

Another way of writing the approximation would be

$$\frac{1}{\sqrt{2}} \leqq x \leqq \sqrt{2}, \qquad z = \frac{x-1}{x+1}$$

$$\log_2 x = c_1 z + c_3 z^3 + c_5 z^5 \pm (.32)\ .10^{-7}.$$

b) Continued fraction approximation, three coefficients:

$$\frac{1}{2} \leqq f \leqq 1, \qquad z = \frac{f - \frac{\sqrt{2}}{2}}{f + \frac{\sqrt{2}}{2}}$$

$$\log_2 f = g_3^*(z) - \frac{1}{2} \pm (.48)\ .10^{-8} = z\left[H_0 + \frac{G_1}{z^2 + H_1}\right] - \frac{1}{2} \pm (.48)\ .10^{-8}$$

Table 3:

$$H_0 = 1.29200 \quad 70987$$
$$H_1 = -1.65676 \quad 26301$$
$$G_1 = -2.63985 \quad 77031.$$

c) Use of increments:

The increments of Tables 1′ and 2′ should be added to the coefficients:

$$(2 \log_2 e, \tfrac{2}{3} \log_2 e, \tfrac{2}{5} \log_2 e, \cdots)$$

and

$$(2 \log_{10} e, \tfrac{2}{3} \log_{10} e, \tfrac{2}{5} \log_{10} e, \cdots)$$

respectively.

In our computations we have used the constants

$$2 \log_2 e = 2.88539\ 00817\ 77926\ 8146$$

$$2 \log_{10} e = .86858\ 89638\ 06503\ 6553.$$

IBM Corporation, New York and
College of the City of New York

1. K. Spielberg, *IB CTR, Chebyshev Truncation System*, Writeup of Program, SHARE Distr. n. 1008.

2. K. Spielberg, "The representation of power series in terms of polynomials, rational approximations and continued fractions," *J. Assoc. Comput. Mach.*, v. 8, 1961, p. 613.

3. C. Hastings, Jr., *Approximations for Digital Computers*, Princeton Univ. Press, New Jersey, 1955.

4. E. G. Kogbetliantz, "Generation of elementary functions," in Ralston & Wilf, *Mathematical Methods for Digital Computers*, John Wiley & Sons, Inc., New York, 1959.

5. E. G. Kogbetliantz, "Papers on elementary functions," *IBM J. Res. Develop.* v. 1, n. 2; v. 2, n. 1; v. 2, n. 3; v. 3, n. 2; 1957–1959.

6. K. Spielberg, "Efficient continued fraction approximations to elementary functions," *Math. Comp.*, v. 15, 1961, p. 409.

TECHNICAL NOTES AND SHORT PAPERS

Some Remarks on Modular Arithmetic and Parallel Computation

By H. S. Shapiro

1. Introduction. A question that has been discussed in recent years is that of *parallel computation*. Can a given computation be broken up into independent assignments which may be performed simultaneously? Traditional methods of computation are almost entirely *serial*, with the consequence that one cannot convert extra computing capacity into significantly greater speed. Thus far the only general method which has been proposed for achieving parallelism is the use of "modular arithmetic"—that is, for some collection of relatively prime integers $m_1, \cdots m_k$ one performs the calculations (mod m_i) independently; the final result is then obtained by solving a system of simultaneous congruences. Such a procedure is possible provided that (i) the calculation consists entirely of additions and multiplications of integers* (so that the corresponding calculations (mod m_i) are justified), and (ii) each number sought in the calculation is an integer known *a priori* to lie in an interval of length $\leqq m_1 m_2 \cdots m_k$.

Modular arithmetic, when applicable, has the advantage of being free from round-off errors; moreover, addition and multiplication (mod m) are carry-free. Another feature is that in some types of calculation (for instance, tabulation of the values of a polynomial for equally spaced values of the argument) the calculation (mod m) is much simplified by the *periodic repetition* of the values being calculated. It therefore seems of interest to show how computations of practical importance may be carried out within the limitations (i) and (ii) above. In this note we discuss division, linear equations, and the first boundary value problem from the standpoint of modular arithmetic.

2. Division. Let us consider the problem of finding d binary digits of the quotient $\frac{x}{y}$ where x and y are integers, $0 < x < y$. The most natural approach is to choose an $n > 0$, and let r be the least non-negative residue of 2^n (mod y), then, writing $N = \frac{2^n - r}{y}$, N is an integer easily computed (mod m) if $(m, y) = 1$, and we have

$$\frac{xN}{2^n} \leqq \frac{x}{y},$$

these numbers differing by $\frac{xr}{2^n y} < \frac{x}{2^n}$. Hence, the integer xN, converted to binary

Received May 22, 1961. This report represents results obtained at the Institute of Mathematical Sciences, New York University, under the sponsorship of the Office of Naval Research, United States Navy.

* Division is allowable only when the modulus is relatively prime to the denominator, and the quotient is an integer.

notation, gives d digits of the quotient $\frac{x}{y}$, providing $2^{n-1-d} > x$. Here the approximation is from below; by working instead with

$$N' = \frac{2^n + s}{y}, \qquad s = y - r,$$

we get an approximation from above. The objection to this procedure is that calculation of the residues of $N(\text{mod } m_i)$ is simple only in case $(m_i, y) = 1$, and so each denominator gives rise to a certain set of "forbidden" moduli.

The following division algorithm is free from this defect. Let $b = 2^k$ be the smallest power of 2 not less than y. Then the sequence t_n defined by

$$bt_{n+1} = (b - y)t_n + x$$

converges to $\frac{x}{y}$ (for an arbitrary choice of t_0). Writing $s_n = b^n t_n$, we get

$$(1) \qquad s_{n+1} = (b - y)s_n + b^n x.$$

If s_0 is chosen to be an integer, all the s_n are integers, easily computed and periodic (mod m), for all m without exception. In place of (1), we can use the convergent iteration

$$(2) \qquad s'_{n+1} = -(y - a)s_n' + a^n x$$

where a is the greatest power of 2 not exceeding y. By choosing the better of (1), (2), i.e., (1) or (2) according as $\frac{b-y}{b}$ or $\frac{y-a}{a}$ is smaller (and at least one of these numbers is $<\frac{1}{3}$) we achieve good convergence. The necessary *a priori* estimate of s_n (or s_n') respectively, and the degree of approximation after n iterations, are readily obtained, and from this the magnitude of $M = \Pi m_i$ sufficient for calculation of $\frac{x}{y}$ to the required accuracy is known. Since t_n arises from s_n upon division by 2^{nk}, i.e., shifting of a binary point, the conversion of s_n from modular to binary notation gives the initial digits of $\frac{x}{y}$ directly.

A variation of this division algorithm which gives a simpler recurrence at the expense of an auxiliary calculation is this. Suppose that the (given) binary expansions of x, y are

$$x = a_0 2^p + a_1 2^{p-1} + \cdots a_p$$
$$y = b_0 2^q + b_1 2^{q-1} + \cdots b_q$$

where a_i, b_i are 0 or 1. We may suppose $a_p = b_q = 1$ since multiplication and division by powers of 2 is trivial. Then $\frac{x}{y} = 2^{p-q} f(\frac{1}{2})$, where f denotes the function

$$(3) \qquad f(t) = \frac{a_p + a_{p-1}\,t + \cdots a_0\,t^p}{b_q + b_{q-1}\,t + \cdots b_0\,t^q} = 1 + c_1\,t + c_2\,t^2 + \cdots.$$

It is easy to show that, in the Taylor expansion (3), we have $| c_n | \leq n$th Fibonacci number, so that (3) converges at least* for $|t| < \frac{\sqrt{5}-1}{2} = .618 \cdots$. Moreover, from (3), the integers c_n are readily computed in terms of the a_i, b_i by a recurrence obtained upon cross-multiplying in (3).

But in terms of the c_n the division can be carried out, using the scheme

$$s_{n+1} = 2\, s_n + c_n\,. \tag{4}$$

If s_0 is an integer (say, $s_0 = 0$) the s_n are integers and

$$\lim_{n\to\infty} \frac{s_n}{2^n} = f\left(\frac{1}{2}\right) = 2^{q-p}\,\frac{x}{y}$$

once again, the *a priori* bounds on s_n, and the rate of convergence (which is uniformly rapid with respect to all divisions) is readily obtained.

These algorithms may be adapted, in an obvious way, to any radix.

3. Linear Equations. Given the system of k equations (in vector notation)

$$Ax = b, \qquad A = \| a_{ij} \| \tag{1}$$

where the a_{ij} and b_i are assumed to be integers, the direct adaptation for modular arithmetic is to compute $d = \det A$, and replace (1) by the system

$$Ay = db \tag{2}$$

for the *integer* variables y_i. Operating with moduli m_i such that $(d, m_i) = 1$ (assuming, of course, $d \neq 0$) the solution of (2) (mod m_i) is very simple (say, by Gaussian elimination). Crude *a priori* bounds on the y_i may be obtained (e.g., by Hadamard's determinant inequality) when they are not available from physical or other considerations. However, there is again the objection that this scheme allows "forbidden moduli" which vary with the given problem. Moreover, the solutions x_i are found as quotients, necessitating divisions which are non-trivial. These difficulties disappear when an iterative method is employed. Let us suppose, to keep the discussion simple, that by preliminary transformations (1) has been put into a form where $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$. Let c_i be the integer defined by: (i) c_i has the same sign as a_{ii}; (ii) $|c_i|$ is the least power of 2 not less than $|a_{ii}|$. Let C be the diagonal matrix of the c_i, and $c = \max |c_i|$. Then the system (1) may be rewritten in the form

$$Cx = Dx + b \tag{3}$$

(writing $D = C - A$), and the iterative scheme

$$Cx_{n+1} = Dx_n + b \tag{4}$$

converges to the solution of (1), because of the supposed diagonal-dominance of A. Finally, putting $y_n = c^n x_n$ we get

$$y_{n+1} = (cC^{-1})Dy_n + (cC^{-1})bc^n. \tag{5}$$

[*] The fact that the series (3) converges at least in this circle, i.e., that the polynomial in the denominator cannot vanish in this circle, was discovered in another connection and pointed out to the author by D. J. Newman.

Since cC^{-1} has integer entries, all components of all y_n are integers (if y_0 is so chosen). The iterative scheme (5) is suitable for modular computation; we omit a detailed discussion of rate of convergence and *a priori* bounds.

4. First Boundary-Value Problem. When the equation $\nabla^2 u = 0$, subject to given boundary conditions, is solved numerically by the method of (square) nets, it is customary to use an iterative method of solution which is known to converge at a rate that can be estimated in terms of the geometry of the region. This problem is in principle subsumed in the above discussion, but two factors make it especially simple from the standpoint of modular computation. First, the transformation to integer variables is particularly simple and especially favorable to a solution in binary notation (owing to the special significance of the number 4 in this iteration). Second, the maximum principle gives a good *a priori* bound on the solutions. For, writing the iterative scheme symbolically as

$$u_{n+1} = Au_n + b$$

and setting $4^n u_n = vn$, we get

$$v_{n+1} = 4Av_n + 4^{n+1}b. \tag{1}$$

Suppose that the components of b are integers (i.e., that the given boundary values are integers, which is achieved by shifting a binary point). If, then, the initial values $v_0 = u_0$ are chosen to be integers lying between the smallest and greatest boundary values, all components of the v_n computed from (1) are integers, lying in the range

$$4^n B_1 \leqq v \leqq 4^n B_2$$

where B_1 and B_2 are the min (and max) of the prescribed boundary values.

5. Remarks on Other Iterative Methods. There are many other important iterative methods in numerical analysis, but not all of these seem well adapted to modular computation, because in many cases transformation to integer variables leads to integers that are too large to be computed practically, i.e., an excessively great number of moduli are required. We may illustrate this with a simple example. Suppose we try to solve the equation

$$x^2 + x - \tfrac{1}{8} = 0 \tag{1}$$

by means of the convergent iteration

$$x_{n+1} = -x_n^2 + \tfrac{1}{8}, \qquad x_0 = 0.$$

Letting $2^{2^n} x_n = y_n$ we have

$$y_{n+1} = 2^{2^{n+1}-3} - y_n^2, \qquad y_0 = 0. \tag{2}$$

The numbers y_n defined by (2) are integers, whose initial binary digits coincide with those of the positive root of (1). Moreover, calculation (mod m) of the numbers y_n from (2) is quite trivial. However, since y_n is of the order of 2^{2^n}, even 10 iterations using moduli of the order of 50 would involve us (roughly) in calculating a 1000 binary digit number, by solving a system of 200 simultaneous congruences—a lot of work to solve a quadratic equation. Newton's method would lead to the same difficulties. A feasible method for solving quadratic equations by modular computation can, however, be based upon the Taylor expansion $(1 - 4x)^{-1/2} = \sum_{n=0}^{\infty} \binom{2n}{n} x^n$.

6. Concluding Remarks. Preliminary analysis indicates that parallel computation, using modular arithmetic, is feasible for certain kinds of problems. The parallel computation envisioned here leads very swiftly to a solution encoded "in modular notation." By this is meant, a system of simultaneous congruences, whose solution (in a specified interval), written as a binary number, has as its *initial digits* the binary number which is the goal of the computation. For results of practical value it will probably be necessary, at the very least, to use moduli whose product exceeds 10^{10}. Hence the feasibility of rapid solution of large-scale systems of congruences will determine the timesaving possibilities of the method. Any *a priori* knowledge about the solution, such as might be obtainable from a preliminary rough solution, analog computation, etc., leads to a reduction in the number of necessary moduli, i.e., knowledge of r binary places reduces the product of the m_i needed by a factor 2^{-r}. Again, in such a case as the boundary value problem, where the values of the solution at neighboring net points differ by amounts which can be bounded *a priori*, this fact might lead to a considerable reduction of labor in the "conversion" phase of the problem.

7. Acknowledgment. The author wishes to thank Lt. Col. L. M. Butsch, Jr., Capt. F. M. Brown, and Capt. A. L. Calton, Jr. of the Bionics and Computer Laboratory, Wright Air Development Division, for directing his attention to the area of modular computation, and Dr. D. L. Slotnick of the Westinghouse Electric Corporation, Air Arm Division, for many stimulating discussions on the subject.

Institute of Mathematical Sciences
New York University
New York 3, N. Y.

Permutations with Restricted Position

By Frank Harary

In his book on combinatorial analysis, Riordan [4, p. 163–164] discusses permutations with restricted position and mentions an open question:
"Any restrictions of position may be represented on a square, with the elements to be permuted as column heads and the positions as row heads, by putting a cross at a row-column intersection to mark a restriction. For example, for permutations of four (distinct) elements, the arrays of restrictions for the rencontres and reduced ménage problems mentioned above are

Received July 5, 1961.

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & x & & & \\ 2 & & x & & \\ 3 & & & x & \\ 4 & & & & x \end{array} \qquad \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & x & & & x \\ 2 & x & x & & \\ 3 & & x & x & \\ 4 & & & x & x \end{array}$$

recontres ménages

Since a square of side n has n^2 cells, and a cross may or may not appear in each cell, it is clear that with n elements 2^{n^2} problems are possible (this includes permutations without restriction, for which no cell has a cross). However, many of these are not distinct since, from the enumeration standpoint, the relative rather than the absolute position of the crosses is important; for example, all n^2 problems having just one cross on the board are alike. The exact number of distinct problems, for any n, is not known, but some progress in this direction will appear in this chapter."

In this note, we show that the question has been virtually solved in [2], and shall obtain an explicit formula for P_n, "the exact number of distinct problems, for any n." For we shall see that the chromatically nonisomorphic bicolored graphs with n points of each color, which are enumerated in [2], are in a one-to-one correspondence with the distinct problems involving permutations on n objects with restricted position.

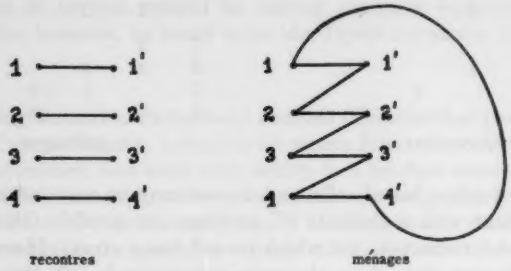
A *binary matrix* is one in which every entry is 0 or 1. Consider the set **M** of all square binary matrices of order n. We say that two matrices A and B in **M** are *equivalent* if B can be obtained from A by the following three operations:

1. Perform any permutation on the rows of A, obtaining A_1
2. Perform any permutation on the columns of A_1, obtaining A_2
3. Either leave the matrix A_2 as it stands or take its transpose, obtaining $A_3 = B$.

Obviously, this is an equivalence relation and it is clear that these three operations are independent. This equivalence relation partitions **M** into equivalence classes. The *number of distinct problems* of permutations on n objects with restricted position is thus the number of equivalence classes of the matrices in **M**. In the above quotation from Riordan [4], the presence of an x in his matrix corresponds to that of a 1 in the associated binary matrix, while a blank space in his matrix becomes a 0 in the binary matrix.

A *graph* consists of a finite collection of *points* together with *lines* joining certain pairs of distinct points. When two points are joined by a line, they are *adjacent*. A graph is said to be *colored* with k colors if each point is assigned one of these colors, any two adjacent points have different colors, and all k colors are used. A *bicolored graph* is one which has been colored with two colors.

To a given restricted permutation problem represented by the binary matrix $A = (a_{ij})$, there corresponds the bicolored graph with $2n$ points $1, 2, \cdots, n, 1', 2', \cdots, n'$ in which point i is joined by a line to point j' if and only if $a_{ij} = 1$. Thus for $n = 4$, the rencontres and reduced ménage problems give the bicolored graphs of Figure 1.



recontres

ménages

FIG. 1

Two graphs are *isomorphic* if there is a one-to-one correspondence between their set of points which preserves adjacency. Two bicolored graphs are *chromatically isomorphic* if there is an isomorphism between them which preserves color, i.e., two points of the first graph have different colors if and only if their image points do. Clearly, chromatic isomorphism is an equivalence relation on bicolored graphs. As an illustration, we show in Figure 2 all bicolored graphs (up to chromatic isomorphism) in which there are two points of each color, together with the corresponding matrices.

LEMMA. *Two square binary matrices are equivalent if and only if the corresponding bicolored graphs are chromatically isomorphic.*

Proof. We translate the three defining operations for equivalence of matrices into graphical terms. Any permutation of the rows of a matrix A in $\mathbf{M}$ corresponds to a renumbering of the n points of the first color in the associated bicolored graph G. A permutation of the columns of A becomes a renumbering of the n points of the second color. Finally, transposing the matrix A amounts to interchanging the two colors assigned to the points of G. Clearly, these three operations serve to characterize chromatic isomorphism.

A formula for the counting polynomial

$$g_{nn}(x) = \sum_{q=0}^{n^2} b_q x^q \tag{1}$$

where b_q is the number of chromatically nonisomorphic bicolored graphs with n points of each color and q lines which have been found in [2]. Let P_n be the number of inequivalent matrices in $\mathbf{M}$, i.e., the number of distinct types of restricted permutation problems (on n objects); then

$$P_n = g_{nn}(1) = b_0 + b_1 + \cdots + b_{n^2}\,. \tag{2}$$

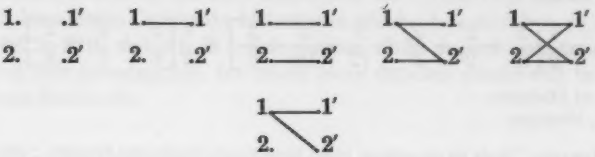
For example, we see from Figure 2 that $g_{22}(x) = 1 + x + 2x^2 + x^3 + x^4$; hence $P_2 = 6$. The number P_n may be found from the cycle index of the "exponentiation group" $S_n^{S_2}$ (where S_n is the symmetric group of degree n) using the enumeration lemma of [1]. This is the same procedure as substituting 1 for x in the formula for $g_{nn}(x)$, which is derived using Pólya's method [3]. To give the result conveniently, we require the following notation:

$(i) = (i_1, i_2, \cdots, i_n)$ denotes a partition of n such that:

$$i_1 + 2i_2 + \cdots + ni_n = n, \tag{3}$$

$$\nu(i) = \prod_{k=1}^{n} k^{i_k} i_k! \tag{4}$$

and $d(r, s)$ is the greatest common divisor of the positive integers r and s.



$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

FIG. 2

We may now state the formula for P_n. Let

$$\alpha(i, j) = \sum_{r,s=1}^{n} i_r j_s d(r, s) \tag{5}$$

$$\beta(j) = \sum_{k \text{ even}} k\left[\frac{j_k}{2} + \binom{j_k}{2}\right] + \sum_{k \text{ odd}} \left[(k+1)\frac{j_k}{2} + k\binom{j_k}{2}\right] + \sum_{r<s} j_r j_s d(r, s). \tag{6}$$

Then

$$P_n = \frac{1}{2} \sum_{(i),(j)} \frac{1}{\nu(i)\nu(j)} 2^{\alpha(i,j)} + \frac{1}{2} \sum_{(j)} \frac{1}{\nu(j)} 2^{\beta(j)} \tag{7}$$

where the first sum is taken over all pairs (i), (j) of partitions of n and the second sum is over all partitions (j) of n.

We illustrate for $n = 3$ whose three partitions π_1, π_2, π_3 are $1 + 1 + 1$, $1 + 2$, and 3. These may be written as the sequences

$$(3, 0, 0), \qquad (1, 1, 0), \qquad (0, 0, 1)$$

respectively. The values of $\alpha(\pi_i, \pi_j)$ for $n = 3$ are given in the matrix:

$$\alpha(\pi_i, \pi_j) \qquad \begin{array}{cc} & j \\ i & \begin{bmatrix} 9 & 6 & 3 \\ 6 & 5 & 2 \\ 3 & 2 & 3 \end{bmatrix} \end{array}$$

while the values of $\nu(\pi_i)$ and $\beta(\pi_i)$ are given in the table:

i	$\nu(\pi_i)$	$\beta(\pi_i)$
1	6	6
2	2	3
3	3	2

Hence we have

$$P_3 = \frac{1}{2}\left[\frac{2^9}{6^2} + \frac{2^5}{2^2} + \frac{2^3}{3^2} + \frac{2^7}{12} + \frac{2^4}{18} + \frac{2^3}{6}\right] + \frac{1}{2}\left[\frac{2^6}{6} + \frac{2^3}{2} + \frac{2^2}{3}\right] = 26.$$

University of Michigan
Ann Arbor, Michigan

1. F. Harary, "Note on an enumeration theorem of Davis and Slepian," *Michigan Math. J.*, v. 3, 1955–56, p. 149–153.
2. F. Harary, "On the number of bicolored graphs," *Pacific J. Math.*, v. 8, 1958, p. 743–755.
3. G. Pólya, "Kombinatorische Anazahlbestimmungen für Gruppen, Graphen und chemische Verbindungen," *Acta Math.*, v. 68, 1937, p. 145–254.
4. J. Riordan, *An Introduction to Combinatorial Analysis*, Wiley & Sons, New York, 1958

Evaluation of the Zeros of Cross-Product Bessel Functions

By L. Jackson Laslett and William Lewish

1. Introduction. There is considerable interest in the zeros of certain cross-product Bessel functions which arise in solving Bessel's equation subject to Dirichlet or Neumann boundary conditions at $r = a, b$,

$$J_n(qa)Y_n(qb) - J_n(qb)Y_n(qa) = 0 \tag{1a}$$

or

$$J_n'(qa)Y_n'(qb) - J_n'(qb)Y_n'(qa) = 0, \tag{1b}$$

because of their well-known application in physical or engineering problems for which the use of cylindrical coordinates is appropriate. In many instances attention may be directed primarily to the zeros of such functions when n is not large because of the interest in the lower-order modes which are possible in the physical problem under consideration, but cases may also arise in which the higher-order modes will warrant attention in order to determine the circumstances in which such possibly unwanted modes may become excited.

Solutions to (1a) and (1b) have been discussed by a number of writers [1], [6], and results presented in the form of algebraic formulas, in tables, or graphically. For application to problems in which $(b - a)/(b + a)$ is small and in which n may

Received February 1, 1961. The work reported in this paper was supported by the U. S. Atomic Energy Commission. It is a pleasure also to acknowledge the assistance provided by Dr. C. L. Hammer in discussing with us the use of the perturbation method for which results are reported in Section 3.

be large, however, it appeared appropriate to make an independent investigation of the initial roots of (1a) and (1b) by study of characteristic solutions of Bessel's equation in the interval $a \leqq r \leqq b$ without explicit reference to the usual Bessel and Neumann functions. Approximate analytic formulas have been obtained from which estimates may be made of the characteristic values, for the case of the first Dirichlet root and for the first two roots subject to the Neumann boundary condition, and an independent numerical determination of the characteristic values and characteristic functions has been made with the CYCLONE electronic digital computer at Iowa State University for cases in which $(b - a)/(b + a)$ was given the values 0.001, 0.01, and 0.1. It is the purpose of the present note to summarize the results of this investigation, for which more detailed results will be available elsewhere (see Section 5).

2. Transformation of Bessel's Equation. It may be noted that, due to the nature of the customary Bessel functions of high order, and in particular because the function J_n remains quite small until its argument is comparable to its order, the lowest characteristic values, q, will be in the neighborhood of n/b for n large. For this reason, and to focus attention on the interval $a \leqq r \leqq b$, it is convenient to define

(2a) $$\eta = \frac{b - a}{b + a},$$

(2b) $$\delta = \eta^2 \left[\left(q \frac{b + a}{2} \right)^2 - n^2 \right],$$

and

(2c) $$x = 2 \frac{r - (b + a)/2}{b - a}.$$

In terms of these quantities,

(3) $$r = \frac{b + a}{2} (1 + \eta x), \quad \text{with} \quad -1 \leqq x \leqq 1,$$

and Bessel's equation assumes the form

(4) $$\frac{d}{dx} \left[(1 + \eta x) \frac{dZ}{dx} \right] + \left[\delta(1 + \eta x) + \frac{2 + \eta x}{1 + \eta x} \cdot \eta^3 n^2 \cdot x \right] Z = 0.$$

The solutions to (4) which are of interest are those for which the Dirichlet boundary condition ($Z = 0$) or, alternatively, the Neumann boundary condition ($dZ/dx = 0$) applies at $x = \pm 1$. When the Dirichlet boundary condition is applied, it may be convenient for some purposes to make the transformation

(5) $$S = (1 + \eta x)^{1/2} Z,$$

in terms of which (4) may be written

(6) $$\frac{d^2 S}{dx^2} + \left[\delta + \frac{(\eta^2/4) + \eta^3 n^2 (2 + \eta x) x}{(1 + \eta x)^2} \right] S = 0,$$

with $S(\pm 1) = 0$.

Physically, it is seen that the quantity η which is introduced here denotes the ratio of the width $(b - a)$ to the mean diameter $(b + a)$ of an annular region. For η only slightly less than unity, the annular region extends substantially from $r = 0$ to $r = b$ and the roots $q\,\dfrac{b+a}{2}$ of (1a) or (1b) may then be expected to become one-half the corresponding roots, μ, of the simpler equations $J_n(\mu) = 0$ or $J_n'(\mu) = 0$, respectively.

For $\eta \ll 1$, the terms in (4) or (6) which contain η, save in some cases those which involve the combination $\eta^3 n^2$, may either be ignored in determining simple analytic formulas for δ or may be treated as a perturbation.

3. Approximate Analytic Formulas. For $\eta \ll 1$, the characteristic values, δ, for (4) or (6) may be obtained by a perturbation method [7] in which the unperturbed equation is taken as simple harmonic, provided n is not too large. In this way we find

(7a) For the first Neumann root: $$\delta \sim \frac{1}{3}\eta^4 n^2 - \frac{8}{15}\eta^6 n^4,$$

(7b) For the first Dirichlet root: $$\delta \sim \left(\frac{\pi}{2}\right)^2 - \frac{\eta^2}{4} + \left(1 - \frac{6}{\pi^2}\right)\left(n^2 - \frac{1}{4}\right)\eta^4,$$

(7c) For the second Neumann root: $$\delta \sim \left(\frac{\pi}{2}\right)^2 + \frac{3}{4}\eta^2 + \left(1 + \frac{10}{\pi^2}\right)\eta^4 n^2.$$

The nature of the characteristic solution associated with the first Neumann root is such that it is very nearly constant when $\eta^3 n^2$ is small. In such cases the form of the solution is approximately given by $Z \sim 1 + \eta^3 n^2\left(x - \dfrac{x^3}{3}\right)$. Similarly, the first Dirichlet and second Neumann solutions are respectively of the general character $\cos\dfrac{\pi}{2}x$ or $\sin\dfrac{\pi}{2}x$. The region of applicability of (7a–c) may be considered to be that for which $\eta^3 n^2 \ll 1$; of equal or greater interest, however, are the results for the case $\eta^3 n^2 > 1$, which is discussed below.

In cases for which $\eta^3 n^2$ is not small, but $\eta \ll 1$, it may suffice to replace (4) by

(8) $$\frac{d^2 Z}{dx^2} + [\delta + 2\eta^3 n^2 \cdot x]Z = 0.$$

Solutions of this approximate equation may be written explicitly in terms of Bessel and Neumann functions of order $\frac{1}{3}$. It then follows, moreover, that for $\eta^3 n^2$ at least somewhat greater than unity (*e.g.*, $\eta^3 n^2 > 6$) the solution of interest is substantially

(9) $$Z \sim \begin{cases} \xi^{1/2}\left[J_{1/3}\left(\dfrac{\xi^{3/2}}{3\eta^3 n^2}\right) + J_{-1/3}\left(\dfrac{\xi^{3/2}}{3\eta^3 n^2}\right)\right], & \text{for } \xi \geqq 0, \\ \dfrac{3^{1/2}}{2}\, i^{4/3}\,|\xi|^{1/2} H^{(1)}_{1/3}\left(i\,\dfrac{|\xi|^{3/2}}{3\eta^3 n^2}\right), & \text{for } \xi \leqq 0, \end{cases}$$

where ξ denotes $\delta + 2\eta^3 n^2 x$, since the first Hankel function then becomes sufficiently small at $x = -1$ as to satisfy adequately the boundary condition normally imposed at that point. The characteristic values, δ, may then be estimated by application of the desired boundary condition at $x = 1$, aided by tables of $J_{\pm 1/3}$ and $J_{\pm 2/3}$ [8], [9],

(10a) For the first Neumann root: $\delta \sim -2\eta^3 n^2 + 1.61724\eta^2 n^{4/3}$,

(10b) For the first Dirichlet root: $\delta \sim -2\eta^3 n^2 + 3.71151\eta^2 n^{4/3}$,

(10c) For the second Neumann root: $\delta \sim -2\eta^3 n^2 + 5.15619\eta^2 n^{4/3}$.

The numerical constants which appear in (10a, b) are seen to be, as expected, twice the numerical coefficients given in series developments for the first maximum and first zero of J_n when n is large [9 (Sect. 15.83, p. 521)]; [10 (Sect. VIII.3.6, p. 143)]. Characteristic values for solutions to (8) must necessarily be somewhat less negative than $-2\eta^3 n^2$ in order that the coefficient of Z be positive for *some* values of x in the interval $-1 \leqq x \leqq 1$. For $\eta^3 n^2$ large, the characteristic solutions are relatively large only for values of x near unity, in a region whose width is roughly two or three times $(\eta^3 n^2)^{-1/3}$.

4. Computational Results. The differential equation (4), suitably scaled, was integrated with the CYCLONE digital computer at Iowa State University, using the Runge-Kutta process [11], [12]. Runs were made for several values of n, with η given in turn the values 0.001, 0.01, and 0.1. In each case the value of δ was adjusted, by trial, to give solutions satisfying the desired Dirichlet or Neumann boundary conditions. A larger number of integration steps was employed to traverse the interval $-1 \leqq x \leqq 1$ in cases in which $\eta^3 n^2$ was large, since more rapid changes of the function occur in certain portions of that interval in such cases. The effect of truncation error was found, by tests in which the interval size was halved, to be sufficiently small that use of the finer interval only affected the final value for the function or its derivative (in the Dirichlet or Neumann cases, respectively) by less than 10^{-6} of the maximum value and the consequent error in δ could thus be judged when tabulating the results of the investigation.

The characteristic values δ determined computationally are listed in Table I. By comparing calculated values of δ obtained for (7a–c) and (10a–c) with the values in Table 1, the accuracy of (7a–c) and (10a–c) can be ascertained. See Table VI [13]. Figure 1 depicts the nature of the associated characteristic functions, for $n = 0.01$, for various representative values of n. Since the contribution from δ makes a relatively small change in the characteristic value for the original Bessel equation when n is large, use of (2b) in connection with the values of δ given in Table I should afford accurate characteristic values for q in such cases. In the application to physical problems it is interesting to note from Figure 1 the features mentioned in Section 3, namely that at small n the first Neumann solution does not show a pronounced variation with x and the other characteristic solutions have approximately the form of circular functions, while at large n the characteristic solutions become large only in a small interval near $x = 1$.

5. Availability of Detailed Results. The analytic work of Section 3 is presented in greater detail, and the computational results reproduced directly from the teleprinter output of the CYCLONE, in an Ames Laboratory report [13]. The report

Table 1

Values of δ for the first Neumann eigenvalue (N_1), the first Dirichlet eigenvalue (D_1), and the second Neumann eigenvalue (N_2).

η Root		$\eta = 0.001$			$\eta = 0.01$			$\eta = 0.1$			$\eta = 1.0$*	
	N_1	D_1	N_2	N_1	D_1	N_2	N_1	D_1	N_2	N_1	D_1	N_2
$n = 0$	0	2.4674011	2.46740209	0	2.467376	2.467476	0	2.4648915	2.4749309	(0)	1.445797	2.670493
½	0	2.4674011	2.46740209	0	2.467376	2.467476	0.00000839	2.4649013	2.474989	0.089633	2.2174011	5.049703
1	0	2.4674011	2.46740209	0	2.467376	2.467476	0.0000330	2.4649308	2.4751348	−0.152511	2.670493	6.106071
2	0	2.4674011	2.46740209	0	2.467376	2.467476	0.0001255	2.4650481	2.4757409	−1.667909	2.593654	7.243048
5	0	2.4674011	2.46740209	0	2.467376	2.467476	0.0005019	2.4658332	2.4802714	−14.709967	−5.765268	2.666867
10	0	2.4674011	2.46740209	0	2.467377	2.467478	−0.0019893	2.4681254	2.4994189	−65.361612	−47.614970	−32.367035
20	0	2.4674021	2.46740209	.0000012	2.467378	2.467484	−0.0680405	2.4693426	2.6189112	−276.5774	−238.492	−208.010
30							−0.335704	2.4435640	2.9276509			
40							−0.905569	2.3513268	3.4493650			
50	0	2.4674011	2.46740210	.0000050	2.467386	2.467520	−1.799205	2.143523	4.0678274			
75							−5.418144	0.8024951	5.142091			
100	0	2.4674011	2.46740211	−0.000020	2.467409	2.467720	−11.009312	−2.103247	4.361139			
150							−28.210824	−13.285380	−2.541564			
200	0.00000002	2.4674011	2.46740217	−0.000720	2.467421	2.468957	−53.56333	−31.956122	−16.514506			
500	0.00000004	2.4674015	2.4674026	−0.032070	2.463969	2.498479						
1000	−0.00000020	2.4674014	2.4674046	−0.456025	2.401502	2.835214						
1500				−1.717081	2.130264	3.657827						
2000	−0.00000720	2.4674015	2.4674169	−3.876713	1.441630	4.535002						
2500				−6.921675	0.145494	4.843058						
3000				−10.857817	−1.875544	4.303889						
4000				−21.440133	−8.287472	0.828077						
5000	−0.00032495	2.467367	2.4677163	−35.682279	−17.993822	−5.740178						
10000	−0.005290	2.466738	2.4718184									
20000	−0.082838	2.456331	2.533433									
50000	−2.030430	2.042788	3.787091									
75000	−6.127374	0.515508	4.811675									
80110		0										
100000	−12.476797	−2.763631	3.907135									
130953			0									
150000	−32.066830	−15.395587	−3.890509									
200000	−60.999918	−36.537911	−19.657327									

* From published tables [8].

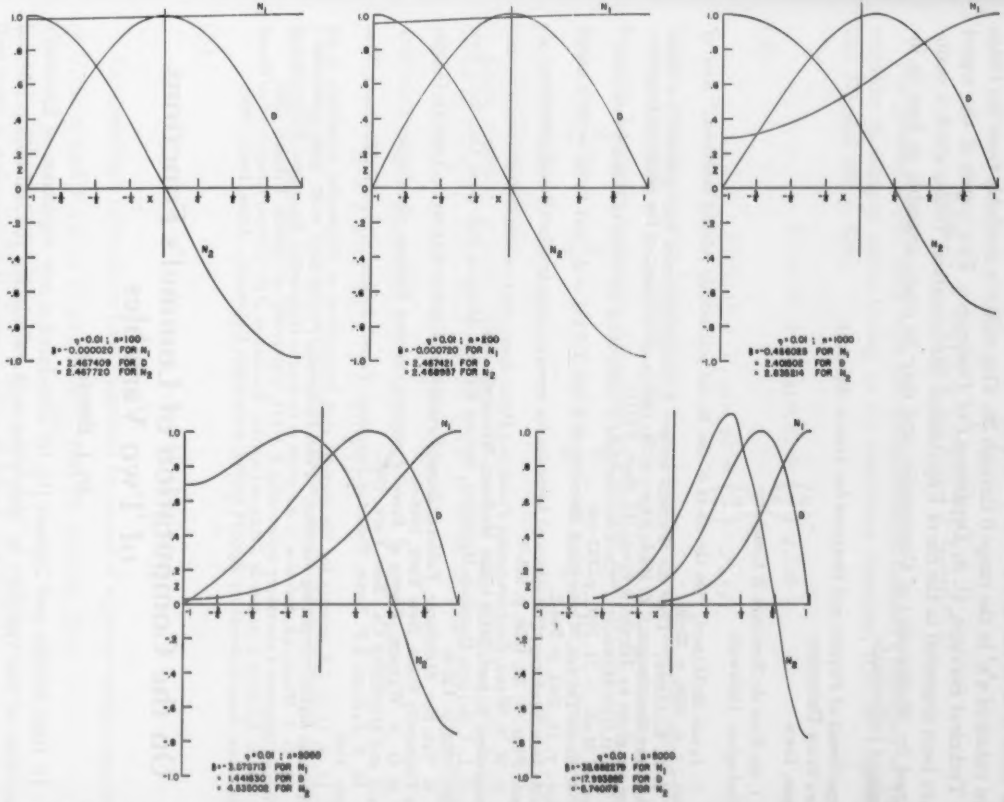


FIG. 1–Characteristic functions for $\eta = .01$ illustrating the effect of n for the first Neumann eigenvalue (N_1), the first Dirichlet eigenvalue (D), and the second Neumann eigenvalue (N_2).

also includes approximate values of $\int_{-1}^{1} Z^2 dx$, suitably normalized with respect to the value of Z or dZ/dx at $x = 0$ and at $x = 1$, for $\eta = 0.0001$ and for representative values of $\eta^2 n^2$ in the range 0 through 20. This report is available from the Office of Technical Services, U. S. Department of Commerce. Two copies of the report have been deposited in the file of Unpublished Mathematical Tables which is maintained by *Mathematics of Computation* and may be made available on loan to interested individuals.

Department of Physics and Institute for Atomic Research
Iowa State University
Ames, Iowa

E. I. du Pont de Nemours & Company
Wilmington, Delaware

1. James McMahon, "On the roots of the Bessel and certain related functions," *Ann. of Math.*, v. 9, 1894, p. 23–30.
2. A. Kalähne, "Über die Wurzeln einiger Zylinderfunktionen und gewisser aus ihnen gebildeter Gleichungen," *Z. Math. Phys.*, v. 54, 1907, p. 55–86. Some of the results of this reference are given by Jahnke & Emde [10 (Sect. VIII)].
3. William Marshall, "On a new method of computing the roots of Bessel's functions," *Ann. of Math.*, v. 11, 1910, p. 153–160.
4. Rohn Truell, "Concerning the roots of $J_n'(x)\, N_n'(kx) - J_n'(kn)\, N_n'(x) = 0$," *J. Appl. Phys.*, v. 14, 1943, p. 350–2.
5. Don Kirkham, "Graphs and formulas for zeros of cross product Bessel functions," *J. Math. Phys.*, v. 36, 1958, p. 371–7.
6. W. N. Wong, *Electromagnetic Fields in a Donut Space*, Midwestern Universities Research Association Report, MURA-555, Madison, Wisconsin, 1960.
7. L. I. Schiff, *Quantum Mechanics*, Second Edition, McGraw-Hill Book Co., Inc., New York, 1955, p. 151–4.
8. Nat. Bur. Standards, *Tables of Bessel Functions of Fractional Order*, v. I and II, Columbia University Press, New York, 1948–9.
9. G. N. Watson, *Theory of Bessel Functions*, Second Edition, Cambridge University Press and Macmillan Co., New York, 1948.
10. E. Jahnke & F. Emde, *Tables of Functions*, Fourth Edition, Dover Publications, New York, 1945.
11. S. Gill, "A process for the step-by-step integration of differential equations in an automatic digital computing machine," *Proc. Cambridge Philos. Soc.*, v. 47, 1951, p. 96–108.
12. D. J. Wheeler, *Solution of a System of Ordinary Differential Equations*, University of Illinois Computer Laboratory subroutine F 1–114, University of Illinois, Urbana, 1953.
13. L. Jackson Laslett & William Lewish, *Evaluation of the Zeros of Cross Product Bessel Functions*, Ames Laboratory Report IS-189, Iowa State University, Ames, Iowa, 1960.

On the Computation of Lommel's Functions of Two Variables

By J. Boersma

In 1942 Zernike and Nijboer [1], [2] introduced a new expansion of Lommel's functions of two variables in connection with calculating th diffraction integral of a circular aperture. In this article it is shown that this expansion is very well suited for the computation of these functions. (The author is much indebted to Dr. Bottema of the Physical Laboratory of the University of Groningen, who drew his attention to this formula.)

Lommel's functions of two variables are defined in the following way (Cf. [3],

Received May 22, 1961.

formulas 16.5 (5) and (6), p. 537, 538),

$$
\begin{aligned}
U_\nu(w, z) &= \sum_{m=0}^{\infty} (-1)^m \left(\frac{w}{z}\right)^{\nu+2m} J_{\nu+2m}(z) \\
V_\nu(w, z) &= \cos\left(\frac{w}{2} + \frac{z^2}{2w} + \frac{\nu\pi}{2}\right) + U_{-\nu+2}(w, z).
\end{aligned}
\tag{1}
$$

The present article deals with the computation of Lommel's functions of two variables of integral order. Owing to the recurrence formulas, (Cf. [3], formulas 16.5 (7) and (8), p. 538),

$$
\begin{aligned}
U_\nu(w, z) + U_{\nu+2}(w, z) &= \left(\frac{w}{z}\right)^{\nu} J_\nu(z) \\
V_\nu(w, z) + V_{\nu+2}(w, z) &= \left(\frac{w}{z}\right)^{-\nu} J_{-\nu}(z),
\end{aligned}
\tag{2}
$$

it is sufficient to compute Lommel's functions for two successive integral values of ν.

The first table of Lommel's functions of two variables of integral order is to be found in Lommel's memoir on diffraction at a circular aperture [4]. Lommel gives tables for $\frac{2}{w}U_1(w, z)$, $\frac{2}{w}\,U_2(w, z)$, and for $\frac{2}{w}\,V_0(w, z)$, $\frac{2}{w}\,V_1(w, z)$ to six decimal places for values of the arguments $w = \pi(\pi)10\pi$, $z = 0(1)12$, and $w = \pi(\pi)8\pi$, $z = 0(1)12$ respectively.

Quite recently, a table [5] by Dekanosidze has been published which gives tables of $U_1(w, z)$, $U_2(w, z)$, $V_1(w, z)$, $V_2(w, z)$ to six decimal places for a somewhat uncommon domain of values of the arguments:

$$
w = 0.5(0.02)1.2(0.05)4(0.1)6.2, \qquad z = w(0.01)4\sqrt{w}
$$

$$
w = 6.3(0.1)10, \qquad z = w(0.01)10.
$$

The tables may also be used outside this domain of values by means of the relations (Cf. [5], formulas (7) and (8))

$$
\begin{aligned}
U_n(w, z) &= (-1)^n V_n\left(\frac{z^2}{w}, z\right) \\
V_n(w, z) &= (-1)^n U_n\left(\frac{z^2}{w}, z\right).
\end{aligned}
\tag{3}
$$

Dekanosidze's tables have been computed by means of a power series expansion of $U_\nu(w, z)$ and $V_\nu(w, z)$ in powers of $\frac{z^2}{2w}$ (Cf. [5], formula (3)),

$$
\begin{aligned}
U_\nu(w, z) &= \sum_{m=0}^{\infty} (-1)^m \frac{1}{m!}\left(\frac{z^2}{2w}\right)^m U_{\nu+m}(w, 0) \\
V_\nu(w, z) &= \sum_{m=0}^{\infty} (-1)^m \frac{1}{m!}\left(\frac{z^2}{2w}\right)^m V_{\nu-m}(w, 0).
\end{aligned}
\tag{4}
$$

When ν is integral, the coefficients of these power series contain a factor of the type $U_n(w, 0)$ and $V_n(w, 0)$, where n is an integer. $U_n(w, 0)$ and $V_n(w, 0)$ are given by [5], formula (4), which contains some printing errors. (Cf. [3], formulas 16.52 (11)–(16), p. 540). The correct formulas are as follows:

$$
\begin{aligned}
U_{2n}(w,0) &= (-1)^n \left[\cos\frac{w}{2} - \sum_{m=0}^{n-1} (-1)^m \frac{\left(\frac{w}{2}\right)^{2m}}{(2m)!} \right] \\
U_{2n+1}(w,0) &= (-1)^n \left[\sin\frac{w}{2} - \sum_{m=0}^{n-1} (-1)^m \frac{\left(\frac{w}{2}\right)^{2m+1}}{(2m+1)!} \right] \\
U_{-n}(w,0) &= \cos\left(\frac{w}{2} + \frac{n\pi}{2}\right) \qquad (5) \\
V_0(w,0) &= 1,\ V_{n+1}(w,0) = 0 \\
V_{-2n}(w,0) &= (-1)^n \sum_{m=0}^{n} (-1)^m \frac{\left(\frac{w}{2}\right)^{2m}}{(2m)!} \\
V_{-2n-1}(w,0) &= (-1)^n \sum_{m=0}^{n} (-1)^m \frac{\left(\frac{w}{2}\right)^{2m+1}}{(2m+1)!}.
\end{aligned}
$$

The computation of expressions (5) for not too small values of w may suffer from loss of digits owing to the alternating character of the series. This same objection arises in computing the alternating series (4) when z is not small.

We now turn to Zernike's method. Here Lommel's functions of two variables are expanded in products of Bessel functions

$$
\begin{aligned}
U_1(w,z) + iU_2(w,z) &= we^{\frac{1}{4}iw} \int_0^1 J_0(zt) e^{-\frac{1}{4}iwt^2} t\,dt \\
&= we^{\frac{1}{4}iw} \sum_{n=0}^{\infty} i^n (2n+1) \sqrt{\frac{2\pi}{w}}\, J_{n+\frac{1}{2}}\left(\frac{w}{4}\right) \frac{J_{2n+1}(z)}{z}.
\end{aligned} \qquad (6)
$$

The essential advantage of this expansion is the following. In (6) all terms of the infinite sum have an absolute value smaller than 1 for all real values of w and z, so, contrary to Dekanosidze's method, there is no danger of loss of digits. This is readily proved by applying the recurrence formula for Bessel functions (Cf. [3], formula 3.2(1), p. 45)

$$2(2n+1)\,\frac{J_{2n+1}(z)}{z} = J_{2n}(z) + J_{2n+2}(z)$$

hence

$$\left| (2n+1)\,\frac{J_{2n+1}(z)}{z} \right| \leq \frac{1}{2} \left| J_{2n}(z) \right| + \frac{1}{2} \left| J_{2n+2}(z) \right| \leq 1.$$

Similarly for $J_{n+\frac{1}{2}}(x)$ the following integral representation is valid:

$$J_{n+\frac{1}{2}}(x) = (-i)^n \sqrt{\frac{x}{2\pi}} \int_{-1}^{+1} e^{ixt} P_n(t)\,dt,$$

(Cf. [3], formula 3.32(2), p. 50) which may be estimated by

$$|J_{n+\frac{1}{2}}(x)| \leqq 2\sqrt{\frac{x}{2\pi}} = \sqrt{\frac{2x}{\pi}};$$

hence

$$\left|\sqrt{\frac{2\pi}{w}}\, J_{n+\frac{1}{2}}\left(\frac{w}{4}\right)\right| \leqq \sqrt{\frac{2\pi}{w}}\sqrt{\frac{w}{2\pi}} = 1.$$

Another advantage is that U_1 and U_2 are calculated simultaneously because each of them is found by adding alternate terms of one single expansion.

The Bessel functions of odd and semi-odd order which are required in equation (6) may be computed very suitably by means of the recurrence technique developed by Goldstein and Thaler [6]. When computing a table of Lommel's functions on an electronic computer, it is possible to store these sequences of Bessel functions, after which various values of w and z may be combined to give $U_1(w, z)$ and $U_2(w, z)$.

The method may still be used, even for large values of w and z, though in that case a rather large sequence of Bessel functions must be computed.

A comparison of the two methods has been made for the case $w = 20$, $z = 20$. When Dekanosidze's method was followed for both functions $U_1(w, z)$ and $U_2(w, z)$, a total of 31 terms of the series in equation (4) had to be taken into account, each term being computed to twelve digits in order to obtain an accuracy of four decimal places (hence a loss of eight digits). When the method described here was followed, 11 terms were already sufficient to give simultaneously $U_1(w, z)$ and $U_2(w, z)$ with the same accuracy without any loss of digits.

Finally, the method described here has been used to recompute Lommel's original tables [4] (the functions $V_0(w, z)$ and $V_1(w, z)$ have been computed by equation (1)), the results being given below. In these tables, the decimals which deviate from Lommel's values have been italicized. Besides that, all values of $\frac{2}{w} V_1(w, z)$ differ by a factor -1 from Lommel's values because the definition of $V_n(w, z)$, as used in the present article and in [3], differs by a factor $(-1)^n$ from Lommel's original definition. (See the footnote at the bottom of [3], p. 537.)

Mathematical Institute
University of Groningen
The Netherlands

1. B. R. A. Nijboer, *The Diffraction Theory of Aberrations*, Thesis, Groningen, 1942, p. 42–43.

2. F. Zernike & B. R. A. Nijboer, "Théorie de la diffraction des aberrations," p. 227–235 of *La Théorie des Images Optiques*, Colloque sur la théorie des images optiques 1947, published by La Revue d'Optique, Paris, 1949.

3. G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Second Edition, Cambridge University Press, 1948.

4. E. Lommel, "Die Beugungserscheinungen einer kreisrunden Öffnung und eines kreisrunden Schirmchens," *Abh. der Math. Phys. Classe der Kön. Bayer. Akad. der Wiss.*, Bd. XV, 1886, p. 229–328.

5. E. N. Dekanosidze, *Tables of Lommel's Functions of Two Variables*, Pergamon Press, New York, 1960.

6. M. Goldstein & R. M. Thaler, "Recurrence techniques for the calculation of Bessel functions," *MTAC*, v. 13, 1959, p. 102–108.

TABLE OF LOMMEL'S FUNCTIONS OF TWO VARIABLES

$w = \pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	+0.636620	+0.636620	+0.636620	0
1	+0.539802	+0.580*638*	+0.4797*44*	−0.088772
2	+0.298890	+0.433460	+0.055002	−0.213022
3	+0.032*426*	+0.247396	−0.383136	−0.055*402*
4	−0.142282	+0.081868	−0.275023	+0.384893
5	−0.173625	−0.022258	+0.450640	+0.252583
6	−0.094496	−0.056474	+0.278235	−0.636026
7	+0.011819	−0.041090	−0.676734	−0.023425
8	+0.070711	−0.008787	+0.430319	+0.531656
9	+0.059110	+0.013939	−0.189447	−0.544147
10	+0.007050	+0.017334	+0.148505	+0.630010
11	−0.035803	+0.007209	−0.245499	−0.620117
12	−0.039518	−0.00430*2*	+0.50*4943*	+0.342522

$w = 2\pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	0	+0.636620	+0.318310	0
1	−0.047572	+0.559947	+0.242644	−0.022268
2	−0.156737	+0.362318	+0.059998	−0.057118
3	−0.250135	+0.124194	−0.115909	−0.041158
4	−0.259807	−0.066375	−0.159699	+0.044515
5	−0.172632	−0.155073	−0.025675	+0.118189
6	−0.036806	−0.141277	+0.164916	+0.050182
7	+0.073194	−0.0682*76*	+0.1629*49*	−0.145567
8	+0.106459	+0.007067	−0.111169	−0.189077
9	+0.06512*5*	+0.04*5577*	−0.26*8535*	+0.11665*1*
10	−0.0046*75*	+0.04071*4*	+0.07368*4*	+0.3119*23*
11	−0.049843	+0.01219*7*	+0.32390*0*	−0.114359
12	−0.046338	−0.013333	−0.155667	−0.331053

$w = 3\pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	−0.212207	+0.212207	+0.212207	0
1	−0.221811	+0.1508*53*	+0.1621*06*	−0.009903
2	−0.233157	−0.000541	+0.044154	−0.025710
3	−0.209291	−0.162866	−0.065351	−0.020817
4	−0.127691	−0.255583	−0.096321	+0.012549
5	−0.005131	−0.240383	−0.034488	+0.046240
6	+0.107300	−0.137968	+0.062157	+0.036719
7	+0.156654	−0.0101*33*	+0.0993*45*	−0.025132
8	+0.124279	+0.079105	+0.025841	−0.081135
9	+0.038751	+0.098281	−0.095896	−0.046848
10	−0.043777	+0.05*9355*	−0.11*6648*	+0.074775
11	−0.077056	+0.00189*8*	+0.03068*3*	+0.133189
12	−0.05282*5*	−0.035554	+0.171787	−0.00764*6*

Table of Lommel's Functions of Two Variables

$w = 4\pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	0	0	+0.159155	0
1	+0.000759	−0.037360	+0.121669	−0.005572
2	+0.010698	−0.122929	+0.034215	−0.014526
3	+0.043564	−0.194789	−0.045730	−0.012219
4	+0.100101	−0.196607	−0.068629	+0.005486
5	+0.157469	−0.114657	−0.027959	+0.024001
6	+0.179329	+0.013349	+0.035306	+0.021696
7	+0.141278	+0.122492	+0.063628	−0.006592
8	+0.052246	+0.160931	+0.029137	−0.036976
9	−0.046242	+0.119653	−0.038977	−0.033318
10	−0.104762	+0.033666	−0.072886	+0.013463
11	−0.097782	−0.043578	−0.027365	+0.060545
12	−0.039653	−0.073719	+0.061663	+0.044025

$w = 5\pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	+0.127324	+0.127324	+0.127324	0
1	+0.123693	+0.101421	+0.097369	−0.003566
2	+0.116978	+0.043947	+0.027779	−0.009316
3	+0.114163	+0.000633	−0.035346	−0.007972
4	+0.114193	+0.008654	−0.053424	+0.003028
5	+0.103761	+0.068242	−0.022719	+0.014668
6	+0.066399	+0.140571	+0.024568	+0.013915
7	−0.000892	+0.173623	+0.046307	−0.002302
8	−0.077848	+0.137780	+0.024054	−0.020595
9	−0.128770	+0.046262	−0.021724	−0.021117
10	−0.125074	−0.053370	−0.048087	+0.002141
11	−0.066712	−0.110067	−0.027077	+0.029850
12	+0.014393	−0.100916	+0.025355	+0.030738

$w = 6\pi$

z	$\frac{2}{w} U_1(w, z)$	$\frac{2}{w} U_2(w, z)$	$\frac{2}{w} V_0(w, z)$	$\frac{2}{w} V_1(w, z)$
0	0	+0.212207	+0.106103	0
1	−0.005291	+0.187222	+0.081156	−0.002477
2	−0.017713	+0.128841	+0.023335	−0.006476
3	−0.030684	+0.074204	−0.028890	−0.005594
4	−0.041775	+0.052871	−0.043819	+0.001917
5	−0.055411	+0.064625	−0.018991	+0.009906
6	−0.076993	+0.080250	+0.018958	+0.009616
7	−0.103193	+0.064885	+0.036479	−0.000963
8	−0.118371	+0.006400	+0.019824	−0.013120
9	−0.103083	−0.072681	−0.014730	−0.014204
10	−0.050141	−0.129002	−0.035334	−0.000298
11	+0.024503	−0.128183	−0.022325	+0.017291
12	+0.086795	−0.069153	+0.013475	+0.020232

Table of Lommel's Functions of Two Variables

$w = 7\pi$

z	$\frac{2}{w}\,U_1(w, z)$	$\frac{2}{w}\,U_2(w, z)$	$\frac{2}{w}\,V_0(w, z)$	$\frac{2}{w}\,V_1(w, z)$
0	−0.090946	+0.090946	+0.090946	0
1	−0.092742	+0.067502	+0.069570	−0.001820
2	−0.095331	+0.011836	+0.020097	−0.004761
3	−0.0931*84*	−0.042950	−0.024469	−0.0041*36*
4	−0.083669	−0.069547	−0.037187	+0.001326
5	−0.069496	−0.065233	−0.016277	+0.007149
6	−0.055115	−0.050886	+0.015516	+0.007029
7	−0.0405*59*	−0.051440	+0.030185	−0.0004*52*
8	−0.019618	−0.073600	+0.016738	−0.009122
9	+0.014183	−0.098796	−0.011166	−0.010150
10	+0.057961	−0.097344	−0.027951	−0.000826
11	+0.095375	−0.053091	−0.018474	+0.011275
12	+0.104159	+0.020684	+0.008673	+0.014010

$w = 8\pi$

z	$\frac{2}{w}\,U_1(w, z)$	$\frac{2}{w}\,U_2(w, z)$	$\frac{2}{w}\,V_0(w, z)$	$\frac{2}{w}\,V_1(w, z)$
0	0	0	+0.079578	0
1	+0.000190	−0.018684	+0.060878	−0.001393
2	+0.002679	−0.061687	+0.017639	−0.003647
3	+0.010993	−0.099549	−0.021243	−0.003179
4	+0.025878	−0.107904	−0.032324	+0.000974
5	+0.0433*75*	−0.084168	−0.014231	+0.0054*08*
6	+0.057603	−0.046848	+0.013178	+0.005359
7	+0.065631	−0.018864	+0.025804	−0.000228
8	+0.069350	−0.008873	+0.014458	−0.006731
9	+0.071904	−0.005808	−0.009042	−0.007608
10	+0.071829	+0.009152	−0.023198	−0.000876
11	+0.061297	+0.043412	−0.015655	+0.007971
12	+0.031978	+0.082866	+0.006317	+0.010231

	$w = 9\pi$		$w = 10\pi$	
z	$\frac{2}{w}\,U_1(w, z)$	$\frac{2}{w}\,U_2(w, z)$	$\frac{2}{w}\,U_1(w, z)$	$\frac{2}{w}\,U_2(w, z)$
0	+0.070736	+0.070736	0	+0.127324
1	+0.069624	+0.055367	−0.001905	+0.112361
2	+0.067676	+0.020712	−0.006385	+0.077695
3	+0.067323	−0.007570	−0.011132	+0.046173
4	+0.0686*69*	−0.0088*52*	−0.015445	+0.035955
5	+0.068172	+0.017625	−0.021256	+0.047323
6	+0.061099	+0.053530	−0.031103	+0.063679
7	+0.04568*1*	+0.0764*75*	−0.044829	+0.065343
8	+0.024884	+0.076750	−0.0583*21*	+0.044755
9	+0.003842	+0.062426	−0.065889	+0.011066
10	−0.014690	+0.049474	−0.064357	−0.018761
11	−0.032149	+0.046026	−0.055052	−0.034099
12	−0.050776	+0.044629	−0.041669	−0.037907

Additions to Cunningham's Factor Table of $n^4 + 1$

By A. Gloden

This note is the fulfillment of a plan to present in a readily accessible and concise form a complete list of additions to the factor tables of $n^4 + 1$ published by Cunningham [1], which give the prime factors (with certain omissions herein supplied) of all such integers not exceeding $1001^4 + 1$. Cunningham's factorizations were found with the aid of his tables [1] of solutions of the congruence

$$x^4 + 1 \equiv 0 \pmod{p}$$

for $p < 10^5$.

The subsequent tables of S. Hoppenot [2], A. Delfeld [3], and the writer [4] have provided an extension of these congruence tables to include all admissible primes between 10^5 and 10^6.

The factorizations presented in the present note have been extracted from a number of sources. The data corresponding to even values of $n \leqq 442$ and to odd values of $n \leqq 523$ have been published previously by M. Kraitchik [5] and N. G. W. H. Beeger [6]. The remaining data have appeared in a series of papers by the writer [7].

In Cunningham's table of factors of $n^4 + 1$ for $n = 2(2)1000$ there appear 97 incomplete entries. Of these, 66 are now identified as primes, corresponding to the following values of n:

320 328 334 340 352 364 374 414 430 436
442 466 472 476 488 492 494 498 504 516
526 540 550 554 556 566 568 582 584 600
616 624 628 656 690 702 710 730 732 738
742 748 758 760 768 772 778 786 788 798
800 810 856 874 894 912 914 928 930 936
952 962 966 986 992 996

Of the remaining 31 incomplete entries, 14 correspond to primes of the form

$$(n^4 + 1)/17,$$

namely, when n = 648, 678, 682, 706, 746, 784, 790, 818, 842, 876, 882, 892, 954, 988.

Furthermore, $(n^4 + 1)/41$ is a prime when n = 888, 946, and 998. Thus, there remain 14 omissions to be considered in Cunningham's table, for even values of n. These factorizations are now given *in extenso*.

Received May 1, 1961.

n	$n^4 + 1$
426	129553·254209
598	203569·628193
640	174289·962609
698	189017·1255801
714	216841·1198537
820	626929·721169
828	176041·2669977
844	246289·2060273
850	170873·3054937
880	290737·2062673
924	158993·4584689
938	809273·956569
980	780049·1182449
982	137593·6758489

In the companion table of factors of $n^4 + 1$, for $n = 1(2)1001$, there appear 82 incomplete entries, of which 68 have now been shown to correspond to primes of the form $(n^4 + 1)/2$. The related values of n are herewith listed:

403
405
415
419
431
445
449
453
455
463

471
477
479
487
503
505
513
517
523
537

539
543
551
561
567
573
579
605
607
613

623
639
643
649
657
677
681
701
703
713

719
721
725
745
761
769
795
805
811
819

821
829
833
843
845
855
857
879
883
891

895
913
917
919
931
963
965
997

Moreover, $(n^4 + 1)/2 \cdot 17$ is prime for $n = 801, 859, 865, 869,$ and 961; $(n^4 + 1)/2 \cdot 41$ is prime for $n = 957$ and 981. In addition to these entries, it is now known that $(n^4 + 1)/2 \cdot 17^2$ is prime when $n = 1001$.

Consequently, there remain only six entries to be considered, and for these the complete factorizations of $(n^4 + 1)/2$ are as follows:

n	$(n_4 + 1)/2$
565	157217·324089
595	137321·456353
685	147377·746969
889	505777·617473
893	17·104233·179441
941	132961·2948521

In conclusion, I should like to state that this paper was prepared as the result of a suggestion made to me by Dr. J. W. Wrench, Jr. that I consolidate my results and those of other researchers which complement the factorizations of $n^4 + 1$ published by Cunningham.

11 Rue Jean Jaurès
Luxembourg

1. A. J. C. Cunningham, *Binomial Factorisations*, v. I and IV, Francis Hodgson, London, 1923 (especially v. I, p. 113–119).

2. S. Hoppenot, *Tables des Solutions de la Congruence* $x^4 \equiv -1 (mod\ N)$ *pour* $100000 < N < 200000$, Librairie du *Sphinx*, Brussels, 1935.

3. A. Delfeld, "Table des solutions de la congruence $X^4 + 1 \equiv 0 (\text{mod } p)$pour $300000 < p < 350000$," Institut Grand-Ducal de Luxembourg, Section des Sciences, *Archives*, v. 16, 1946, p. 65–70.

4. A. Gloden, "Table des solutions de la congruence $X^4 + 1 \equiv 0 (\text{mod } p)$pour $2 \cdot 10^5 < p < 3 \cdot 10^5$," *Mathematica* (Rumania), v. 21, 1945; *Table des solutions de la congruence* $x^4 + 1 \equiv 0 (mod\ p)$ *pour* $350000 < p < 500000$, Centre de Documentation Universitaire, Paris, 1946; *Table des solutions de la congruence* $x^4 + 1 \equiv 0 (mod\ p)$ *pour* $500000 < p < 600000$, Luxembourg, author, rue Jean Jaurès 11, 1947; *Table des solutions de la congruence* $x^4 + 1 \equiv 0 (mod\ p)$ *pour* $600000 < p < 800000$, Luxembovrg, published by the author, 1952; *Table des solutions de la congruence* $x^4 + 1 \equiv 0 (mod\ p)$ *pour* $800000 < p < 1000000$, Luxembourg, published by the author, 1959.

5. M. Kraitchik, *Recherches sur la Théorie des Nombres*, v. 2, Gauthier-Villars, Paris, 1929, p. 116–117.

6. N. G. W. H. Beeger, *Additions and corrections to "Binomial Factorisations" by Lt. Col. A. J. C. Cunningham*, Amsterdam, 1933.

7. A. Gloden, "Compléments aux tables de factorisation de Cunningham," *Mathesis*, v. 55, 1945–46, p. 254–256; *ibid.*, v. 61, 1952, p. 49–50, 101, 305–306; v. 68, 1959, p. 172. See also *Intermédiaire des Recherches Mathématiques*, v. 4, 1948, p. 39.

On the Generation of All Possible Stepwise Combinations

By Gary Lotto

Conventionally, when all possible combinations of all possible subset sizes from a set of n are desired, a binary count is performed. Associating the units digit with the number 1, the two's digit with the number 2, the four's digit with the number 3, etc., the binary count 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, etc., becomes associated with the combinations 1, 2, 12, 3, 13, 23, 123, 4, etc. This is useful in such procedures as the analysis of variance.

The above order of combinations requires that, when computing on data from one combination to the next, either (a) the calculation starts anew, or (b) if algorithms exist for generating a new function from the old one by single steps of either including or deleting a number from the combination, more than one step may be required. For example, we may go from the combination "2" to the combination "12" by "including 1." But going from "12" to "3" requires "deleting 1, deleting 2, and including 3."

Given, then, that a problem may be solved for some combination of k elements from the solution for the superset of $(k + 1)$ elements or the subset of $(k - 1)$ elements, is there an algorithm for generating all possible combinations which goes through the fewest recursions?

Received May 18, 1961.

TABLE 1

i	$A(i)$	$B(i)$	$C(i)$
1	00001	+1	1
2	00010	+2	1 2
3	00011	−1	2
4	00100	+3	2 3
5	00101	+1	1 2 3
6	00110	−2	1 3
7	00111	−1	3
8	01000	+4	3 4
9	01001	+1	1 3 4
10	01010	+2	1 2 3 4
11	01011	−1	2 3 4
12	01100	−3	2 4
13	01101	+1	1 2 4
14	01110	−2	1 4
15	01111	−1	4
16	10000	+5	4 5

The author has used the following algorithm to generate all combinations of independent variables in a multiple regression problem:

(1) For each step, carry the cycle number i of the combination which is to be generated.

(2) Divide i by 2, then the quotient by 2, etc., until the remainder is not 0. The number of divisions performed is k, the number to be included or deleted.

(3) Divide the quotient of the last division in (2) by 2. If the remainder is 0, include. If the remainder is 1, delete.

The algorithm is equivalent to inspecting the lowest non-zero bit in the binary representation of i. If this is the kth bit (counting from the right), the number k is to be included or deleted. The $(k + 1)$st bit instructs inclusion or deletion: if 0, include; if 1, delete.

Define $A(i)$ as the binary representation of i, $B(i)$ as $+k$ if the number k is to be included, or $-k$ if k is to be deleted on cycle i, and $C(i)$ as the resultant combination. Table 1 gives the first 16 values of i and these functions.

Given combinations 1 through $(2^{k-1} - 1)$, all combinations of $(k - 1)$ elements, the additional combinations which must be generated in order to produce all combinations of k elements are reproductions of the first $(2^{k-1} - 1)$ combinations, to each of which has been added the kth element, plus the combination of element k alone (in effect, a reproduction of the zero combination, plus element k).

The algorithm produces these combinations by: (1) including k on the 2^{k-1}st cycle, and not deleting it before the 2^kth cycle, and (2) reproducing the $B(i)$'s in reverse order with opposite sign $(B(2^{k-1} + c) = -B(2^{k-1} - c))$, thus on each cycle deleting from the combination that which we, $2c$ cycles before, included into it, or including that which we, $2c$ cycles before, deleted from it, until the (2^{k-1})st combination, which corresponds to the empty set plus element k.

Proof of (1). Since the binary representation of 2^{k-1} is a 1 bit followed by $(k - 1)$ zeros, the kth element is included on cycle 2^{k-1}. The kth element will remain until the binary number 11 followed by $(k - 1)$ zeros appears. This will be on cycle number $(2^k + 2^{k-1}) > (2^k - 1)$. Thus, all combinations from 2^{k-1} through $(2^k - 1)$ will include the kth element.

Proof of (2). Since $(2^{k-1} + c) + (2^{k-1} - c) = 2^k$, the binary representations of $(2^{k-1} + c)$ and $(2^{k-1} - c)$ correspond in all their low-order zeros, and the low-order 1, in which they also correspond. The bit above the 1 must differ in the two numbers, due to the binary carry. Thus, $B(2^{k-1} + c) = -B(2^{k-1} - c)$.

To complete the proof by induction, we may note, by Table 1, that the algorithm has generated all combinations for $k \leqq 4$.

University of Pittsburgh, and
American Institute for Research
Pittsburgh, Pa.

Generation of Permutations by Addition

By John R. Howell

1. Introduction. Suppose one wishes to generate the $k!$ permutations of k distinct marks. Representing these k marks by $0, 1, 2, \cdots, (k - 1)$ written side by side to form the "digits" of a base k integer, then the repeated addition of 1 will generate integers whose "digits" represent permutations of k marks. Many numbers are also generated which are not permutations. D. H. Lehmer [2] states that this so-called addition method can be made more efficient by adding more than 1 to each successive integer.

2. Method. In this note, we show that the correct number greater than 1 to add to this integer is a multiple of $(k - 1)$ radix k.

Lemma 1. *The arithmetic difference radix k between an integer composed of mutually unlike digits and another integer composed of a permutation of the same digits is a multiple of $(k - 1)$.*

Considering the process of "casting out nines," it is obvious that the two integers are congruent mod $(k - 1)$. Hence, their difference is zero mod $(k - 1)$.

The method seems to have two advantages. First, one can generate all $k!$ permutations in lexicographic order. Second, all permutations "between" two given permutations can be obtained. The process can be made to be cyclic if upon obtaining $(k - 1), \cdots, 0$ one takes the next permutation to be $0, 1, \cdots, (k - 1)$.

3. Example. Suppose we wish to generate the 4! permutations of 4 marks. Representing these 4 marks by 0, 1, 2 and 3, we add 3 radix 4 to 0123 to get 0132. Continuing this process we get the 4! permutations desired. The array below shows

Received September 11, 1961.

the first 16 numbers generated by this process. An asterisk marks each integer whose digits represent a required permutation. The other integers were rejected because of the occurrence of repeated digits.

Sequence	*Integer*	*Sequence*	*Integer*
1	0123*	9	0303
2	0132*	10	0312*
3	0201	11	0321*
4	0210	12	0330
5	0213*	13	0333
6	0222	14	1002
7	0231*	15	1011
8	0300	16	1020

4. Adaptation to a Computer. In a computer such as the IBM 7090 where convert instructions are available it is easy to do radix k arithmetic. Otherwise one could simulate the process by adding 9 digit-wise and testing the resulting sum for having unique digits each one of which is one of the original k digits.

5. Acknowledgments. This method was developed when the author was with the Statistics Department, Agricultural Experiment Station, University of Florida, Gainesville, Florida, in connection with the problem of obtaining a particular arrangement of the rows of a Latin square. He wishes to thank Mark Robinson of Martin Marietta Corp. for suggestions concerning the writing of the manuscript.

Computer Applications Department
Martin Marietta Corporation
Orlando, Florida

1. C. B. Tompkins, "Machine attacks on problems whose variables are permutations," *Proceedings of Symposia in Applied Mathematics*, v. VI, *Numerical Analysis*, McGraw-Hill, New York, 1956, p. 195–211.

2. D. H. Lehmer, "Teaching combinatorial tricks to a computer," Proceedings of Symposia in Applied Mathematics, v. X, *Combinatorial Analysis*, American Mathematical Society, Providence, R. I., 1960, p. 179–193.

3. Mark B. Wells, "Generation of permutations by transposition," *Math. Comp.* v. 15, 1961, p. 192–195.

Multiple Quadrature with Central Differences on One Line

By Herbert E. Salzer

Abstract. The coefficients A_{2m}^{n} in the n-fold quadrature formulas for the stepwise integration of (1) $y^{(n)} = f(x, y, y', \cdots, y^{(n-1)})$, at intervals of h, namely, for n even, (2) $\delta^n y_0 = h^n \sum_{m=1}^{10} (1 + A_{2m}^{n}\delta^{2m})f_0 + \cdots$, for n odd, (3) $\mu\delta^n y_0 = h^n \sum_{m=1}^{10} (1 + A_{2m}^{n}\delta^{2m})f_0 + \cdots$, are tabulated exactly for $n = 1(1)6$, $m = 1(1)10$. They were calculated from the well-known symbolic formulas (4) $\delta^n y = (\delta/D)^n f$, (5) $(\delta/D)^n = (\delta h/2 \sinh^{-1}(\delta/2))^n$ and (6) $\mu = (1 + \delta^2/4)^{1/2} = 1 + \frac{\delta^2}{8} - \frac{\delta^4}{128} + \frac{\delta^6}{1024} -$

Received March 8, 1961.

$\dfrac{5\delta^8}{32768} + \cdots$. For calculating $y^{(r)}$, replace n by $n - r$ in (2) and (3). Use of (2) and (3) avoids the solution of (1) by simultaneous lower-order systems for $n > 1$, as well as mid-interval tabular arguments, requires only even-order differences, on a single line, and provides great accuracy due to rapid decrease of A^n_{2m} as m increases. However, the integration may be slowed down by the need to estimate and refine iteratively the later values of $y, y', \cdots, y^{(n-1)}$ required in $\delta^{2m} f_0$. Reference to earlier collected formulas of Legendre, Oppolzer, Thiele, Lindow, Salzer, Milne and Buckingham, reveals that Thiele and Buckingham come closest to (2), (3), as their works contain schemes that involve just tabular arguments throughout. For n odd, they give formulas that are based upon the series in δ^{2m} for $(1/\mu)(\delta/D)^n$ instead of $\mu(\delta/D)^n$ as in the present arrangement.

1. Purpose and Scope of Tabulated Formulas. Given a differential equation

$$(1) \qquad y^{(n)} = f(x, y, y', \cdots, y^{(n-1)}),$$

and a sufficient number of starting values at intervals of h, there are very convenient numerical integration formulas for obtaining either $\delta^n y_0$, for n even, or $\mu\delta^n y_0$, for n odd, in terms of just the even-order central differences of $f \equiv f(x, y, y', \cdots, y^{(n-1)})$ at $x = x_0$, denoted by $\delta^{2m} f_0$. This article tabulates the exact values of A^n_{2m}, the coefficients of $\delta^{2m} f_0$, for $n = 1(1)6$, $m = 1(1)10$, in the following numerical integration formulas:

$$(2) \qquad \delta^n y_0 = h^n \sum_{m=1}^{10} (1 + A^n_{2m}\delta^{2m}) f_0 + \cdots, \quad \text{for } n \text{ even, and}$$

$$(3) \qquad \mu\delta^n y_0 = h^n \sum_{m=1}^{10} (1 + A^n_{2m}\delta^{2m}) f_0 + \cdots, \quad \text{for } n \text{ odd.}$$

The computation of A^n_{2m} was based upon the symbolic form of (1), or $D^n y = f$, from which

$$(4) \qquad \delta^n y = (\delta/D)^n f.$$

The well-known operational formula,

$$(5) \qquad (\delta/D)^n = (\delta h/2 \sinh^{-1} (\delta/2))^n,$$

was used to obtain the coefficients of δ^{2m} in the series for $(\delta/D)^n$. For even n, this yielded (2). For odd n, (5) produces integration formulas that express mid-interval values of y in terms of tabular values of f. To obtain (3), which involves tabular values of both y and f, we multiply (5) by μ, giving μ, on the left side, the numerical interpretation of a mean central operator $\frac{1}{2}(E^{1/2} + E^{-1/2})$, and considering μ, on the right side, a symbolic even function of δ according to

$$(6) \qquad \mu = (1 + \delta^2/4)^{1/2} = 1 + \frac{\delta^2}{8} - \frac{\delta^4}{128} + \frac{\delta^6}{1024} - \frac{5\delta^8}{32768} + \cdots.$$

Integration of (1) also requires formulas for the stepwise determination of the derivatives $y^{(r)}$, $r = 1(1)n-1$. By noting that $D^{n-r} y^{(r)} = f$, we can still employ (2) and (3), as well as the same quantities $\delta^{2m} f_0$, merely replacing n by $n - r$.

In the use of (2) and (3) we avoid the widespread practice of breaking up a higher-order equation into a simultaneous first-order system where each equation requires its own set of differences. Also there is no occurrence of formulas involving mid-interval arguments. Among the attractive features of this scheme is the employment of just alternate even-order differences that are on a single line. Besides the concise and economical appearance of (2), (3), the rapid rate of decrease of A^n_{2m} with increasing m is seen to provide high accuracy.

On the dampening side, the user is reminded that the higher-order central differences of $f(x, y, y', \cdots, y^{(n-1)})$ in (2) and (3) involve later values of $y, y', \cdots, y^{(n-1)}$ that must be estimated at first, probably by some kind of extrapolation. Then (2) and (3) might be used in some iterative refining scheme, the details depending upon the particular functional form of $f(x, y, y', \cdots, y^{(n-1)})$, the nature of the problem, and the desired accuracy (all of which is a vast subject in itself).

2. Comparison with Earlier Work. The chief novelty in the present arrangement is the systematic use of the μ-series in terms of δ^{2m} to obtain (3) for any odd n (see also Milne below). Two other authors (Thiele, Buckingham), by employing the series for $1/\mu$ in terms of δ^{2m}, give formulas for odd n that are closely related to (3), requiring just tabular arguments and avoiding the introduction of mid-interval arguments (as is done by Legendre, Oppolzer, Lindow). Presented chronologically, there is the following earlier work.

Legendre [1] gives the symbolic formula for the $(\delta/D)^n$ series in δ^{2m} and the first few coefficients up to $n = 6$.

Oppolzer [2] gives the exact coefficients for (δ/D) and $(\delta/D)^2$ up to δ^{20}. His (δ/D) coefficients checked with those in Salzer [5]. His $(\delta/D)^2$ coefficients checked with A^2_{2m} here, except for his coefficient Q_2^{14} $(= A^2_{16})$ not in lowest terms by a factor of 9.

Thiele [3] gives the exact values of the first five coefficients for D^{-n} and $(1/\mu)D^{-n}$, which is the same as $(\delta/D)^n$ in terms of δ^{2m} and $\mu\delta^{2m}$ up to $m = 5$, for $n = 1(1)5$.

Lindow [4], who gives some central difference formulas up to triple quadrature, also gives the exact values of A^2_{2m}, for $m = 1(1)7$.

Salzer [5] tabulates the coefficients of δ/D, exactly through δ^{20}, then 18D through δ^{50}.

Milne [6] happens to give $2A^1_{2m}$, $m = 1(1)5$, in the first of a series of formulas for $\int_{x_0-rh}^{x_0+rh} f(x)dx$, $r = 1(1)5$, in terms of $\delta^{2m}f_0$.

Salzer [7] gives the coefficients of δ_0^{2m} and δ_1^{2m} obtained by k-fold quadrature of Everett's formula; for $k = 2$, exactly up to $m = 10$, then 16D up to $m = 24$; for $k = 3(1)6$, exactly for $m = 0$ and 8S for $m = 1(1)10$. These differ from the other coefficients in that they occupy two lines for central differences instead of one. They are mentioned here because of their similar purpose and the large extent to which they have been tabulated.

Buckingham [8] gives the coefficients of $(\delta/D)^n$ and $(1/\mu)(\delta/D)^n$, $n = 1(1)4$, through δ^8. As in Thiele [3], this includes an integration scheme involving just tabular arguments for every n. Thus, by expressing $(\delta/D)^n$ for odd n as $\{(1/\mu)(\delta/D)^n\}\mu$, and choosing $x_0 + h/2$ for the argument, Buckingham obtains odd-order central differences of the integral, at mid-intervals, in terms of mean central even-order differences, also at mid-intervals, so that both members involve y and f for just tabular arguments. However, it appears to the author that for n odd there is less total computation involved in using (3) for $\mu(\delta/D)^n$, where the slight extra work of finding $\mu\delta^n y_0$ instead of $\delta^n y_{1/2}$ is more than compensated for by not having to average all the quantities $\delta^{2m} f_0$ and $\delta^{2m} f_1$, as is done in the Buckingham-Thiele procedure which uses $(1/\mu)(\delta/D)^n$ with the mean central differences $\mu\delta^{2m} f_{1/2}$.

3. Integration Formulas for $y^{(n)} = f(x, y, y', \cdots, y^{(n-1)})$

$$\begin{aligned} n = 1: \quad \mu\delta y_0 = h\biggl(1 &+ \frac{\delta^2}{6} - \frac{\delta^4}{180} + \frac{\delta^6}{1512} - \frac{23}{2\ 26800}\delta^8 + \frac{263}{149\ 68800}\delta^{10} \\ &- \frac{1\ 33787}{4\ 08648\ 24000}\delta^{12} + \frac{1\ 57009}{24\ 51889\ 44000}\delta^{14} \\ &- \frac{162\ 15071}{12504\ 63614\ 40000}\delta^{16} + \frac{26894\ 53969}{99\ 78699\ 64291\ 20000}\delta^{18} \\ &- \frac{2\ 68931\ 18531}{4704\ 24411\ 73728\ 00000}\delta^{20}\biggr) f_0 \end{aligned}$$

$$\begin{aligned} n = 2: \quad \delta^2 y_0 = h^2\biggl(1 &+ \frac{\delta^2}{12} - \frac{\delta^4}{240} + \frac{31}{60480}\delta^6 - \frac{289}{36\ 28800}\delta^8 + \frac{317}{228\ 09600}\delta^{10} \\ &- \frac{68\ 03477}{261\ 53487\ 36000}\delta^{12} + \frac{32\ 03699}{627\ 68369\ 66400}\delta^{14} \\ &- \frac{736\ 91749}{71137\ 48561\ 92000}\delta^{16} + \frac{22\ 03877\ 95651}{10218\ 18843\ 43418\ 88000}\delta^{18} \\ &- \frac{15447\ 34732\ 56043}{337\ 20021\ 83332\ 82304\ 00000}\delta^{20}\biggr) f_0 \end{aligned}$$

$$\begin{aligned} n = 3: \quad \mu\delta^3 y_0 = h^3\biggl(1 &+ \frac{\delta^2}{4} + \frac{\delta^4}{240} + \frac{\delta^6}{60480} - \frac{\delta^8}{57600} + \frac{661}{1596\ 67200}\delta^{10} \\ &- \frac{4\ 65967}{52\ 30697\ 47200}\delta^{12} + \frac{3\ 96079}{209\ 22789\ 88800}\delta^{14} \\ &- \frac{95\ 95529}{23712\ 49520\ 64000}\delta^{16} + \frac{1\ 78574\ 25881}{2043\ 63768\ 68683\ 77600}\delta^{18} \\ &- \frac{2143\ 27306\ 64071}{112\ 40007\ 27777\ 60768\ 00000}\delta^{20}\biggr) f_0 \end{aligned}$$

$$\begin{aligned} n = 4: \quad \delta^4 y_0 = h^4\biggl(1 &+ \frac{\delta^2}{6} - \frac{\delta^4}{720} + \frac{\delta^6}{3024} - \frac{41}{7\ 25760}\delta^8 + \frac{491}{479\ 00160}\delta^{10} \\ &- \frac{3\ 41749}{17\ 43565\ 82400}\delta^{12} + \frac{50971}{13\ 07674\ 36800}\delta^{14} \\ &- \frac{1704\ 03199}{2\ 13412\ 45685\ 76000}\delta^{16} + \frac{8\ 55137\ 58923}{5109\ 09421\ 71709\ 44000}\delta^{18} \end{aligned}$$

$$-\frac{1721\ 38184\ 48999}{48\ 17145\ 97618\ 97472\ 00000}\,\delta^{20}\Bigg)f_0$$

$$n = 5:\quad \mu\delta^5 y_0 = h^5\Bigg(1 + \frac{\delta^2}{3} + \frac{\delta^4}{48} - \frac{\delta^6}{6048} + \frac{11}{7\ 25760}\,\delta^8 - \frac{\delta^{10}}{7\ 98336}$$

$$+ \frac{13283}{17\ 43565\ 82400}\,\delta^{12} + \frac{5827}{104\ 61394\ 94400}\,\delta^{14}$$

$$- \frac{9\ 66067}{23712\ 49520\ 64000}\,\delta^{16} + \frac{4757\ 70541}{364\ 93530\ 12264\ 96000}\,\delta^{18}$$

$$- \frac{24\ 19396\ 16497}{6\ 88163\ 71088\ 42496\ 00000}\,\delta^{20}\Bigg)f_0$$

$$n = 6:\quad \delta^6 y_0 = h^6\Bigg(1 + \frac{\delta^2}{4} + \frac{\delta^4}{120} + \frac{\delta^6}{30240} - \frac{\delta^8}{57600} + \frac{31}{76\ 03200}\,\delta^{10}$$

$$- \frac{27257}{3\ 11351\ 04000}\,\delta^{12} + \frac{11581}{6\ 22702\ 08000}\,\delta^{14}$$

$$- \frac{15\ 54079}{3908\ 65305\ 60000}\,\delta^{16} + \frac{1\ 25353\ 54591}{1459\ 74120\ 49059\ 84000}\,\delta^{18}$$

$$- \frac{150\ 48397\ 12643}{8\ 02857\ 66269\ 82912\ 00000}\,\delta^{20}\Bigg)f_0$$

General Dynamics/Astronautics
San Diego, California

1. A. M. Legendre, *Traité des Fonctions Elliptiques*, v. 2, Paris, 1826, Chapter 3, p. 41–60 (For errors, see *MTAC*, v. 5, 1951, p. 27).

2. T. R. Oppolzer, *Lehrbuch zur Bahnbestimmung der Cometen und Planeten*, v. 2, W. Engelmann, Leipzig, 1880, p. 35, 53–54, 545, 596.

3. T. N. Thiele, *Interpolationsrechnung*, B. G. Teubner, Leipzig, 1909, p. 95–97. (Some misprints are noted in *Math. Comp.*, v. 15, 1961, p. 321.)

4. M. Lindow, *Numerische Infinitesimalrechnung*, F. Dümmler, Berlin and Bonn, 1928, p. 170–171.

5. H. E. Salzer, "Coefficients for mid-interval numerical integration with central differences," *Phil. Mag.*, ser. 7, v. 36, 1945, p. 216–218.

6. W. E. Milne, *Numerical Calculus*, Princeton, 1949, p. 196–197.

7. H. E. Salzer, "Coefficients for repeated integration with central differences," *J. Math. Phys.*, v. 28, 1949, p. 54–61.

8. R. A. Buckingham, *Numerical Methods*, Pitman Publishing Corp., New York and London, 1957, p. 150–154. (For errors, see *Math. Comp.*, v. 15, 1961, p. 319.).

New Mersenne Primes

By Alexander Hurwitz

If p is prime, $M_p = 2^p - 1$ is called a Mersenne number. The primes M_{4253} and M_{4423} were discovered by coding the Lucas-Lehmer test for the IBM 7090. These two new primes are the largest prime numbers known; for other large primes see Robinson [4]. The computing was done at the UCLA Computing Facility. This test is described by the following theorem (see Lehmer [1, p. 443–[illegible]]).

Theorem. *If* $S_1 = 4$ *and* $S_{n+1} = S_n^2 - 2$ *then* M_p *is prime if and only if* $S_{p-1} \equiv 0 \pmod{M_p}$.

The test takes about 50 minutes of machine time for $p = 4423$. These results bring the number of known Mersenne primes to 20. The values of p for these twenty primes are listed in Table 1.

If M_p is prime it is of interest to know the sign of the least absolute penultimate residue, that is, whether $S_{p-2} \equiv +2^r \pmod{M_p}$ or $S_{p-2} \equiv -2^r \pmod{M_p}$ where $2r = p + 1$. The Lucas-Lehmer test can also be used with $S_1 = 10$. The various penultimate residues of the known Mersenne primes were computed and the results appear in Table 1 (see Robinson [3]).

In addition to testing the above Mersenne primes each Mersenne number with $p < 5000$ was tested unless a factor of M_p was known. The residues of $S_{p-1} \pmod{M_p}$ are available for checking purposes. The results for $3300 < p < 5000$ are summarized in Table 2. The computer program also found (see [3, p. 844]) that M_{8191} is not prime.

The residue $S_{p-1} \pmod{M_p}$ for $p > 3300$ is output from the computer in a modified octal notation. That is, the residue is stored in the computer in 35 bit binary words and the output is a word by word conversion of the 35 bit words into octal (base 8) numbers. Thus the leading digit of each is quaternary (base 4). For $p < 3300$ the residue was printed in hexadecimal notation (see Robinson [3] and Riesel [2]).

Table 1

p	$S_1 = 4$	$S_1 = 10$	p	$S_1 = 4$	$S_1 = 10$
2			107	−	+
3	+	−	127	+	+
5	+	−	521	−	+
7	−	−	607	−	−
13	+	+	1279	−	−
17	−	+	2203	+	−
19	−	+	2281	−	+
31	+	+	3217	−	+
61	+	+	4253	+	+
89	−	+	4423	−	−

Received November 3, 1961. The preparation of this paper was sponsored by the U. S. Office of Naval Research.

Table 2

p	R	p	R
3301	72013	4241	11012
3307	62061	4253	00000
3313	10050	4259	46007
3331	51270	4261	55632
3343	76415	4283	74774
3371	57040	4339	41356
3373	36120	4349	74465
3389	64705	4357	74271
3413	50261	4363	61114
3461	03241	4397	40174
3463	57665	4409	51070
3467	23046	4421	25131
3469	21765	4423	00000
3547	75574	4481	70216
3559	45350	4493	36053
3583	42507	4519	01571
3607	45062	4523	22235
3617	35431	4567	74267
3631	14530	4583	46556
3637	67413	4591	47243
3643	04606	4621	74601
3671	04031	4643	51444
3673	01626	4651	00707
3691	54715	4663	52442
3697	53743	4673	40333
3709	06427	4679	14305
3739	22413	4703	54013
3769	00747	4721	04420
3821	52075	4729	40137
3833	45453	4733	12774
3847	57652	4783	77350
3877	46507	4789	02364
3881	34503	4799	04305
3889	30737	4817	70020
3919	16520	4831	33213
3943	33442	4877	75412
4007	17770	4889	24410
4027	60265	4909	61113
4049	31260	4937	26525
4051	63236	4951	22271
4091	55650	4973	03354
4093	26670	4987	72275
4111	20437		
4133	66046	8191	03624
4157	43640		
4159	62544		
4177	16076		
4201	53211		
4219	51756		
4231	51457		

The five least significant octal digits of the residue appear in Table 2 for each $p > 3300$ tested. If p $(3300 < p < 5000)$ is omitted from Table 2 a factor of $2^p - 1$ is known. Some of these factors are not yet published but were communicated to the author by John Brillhart.

My thanks to the Computing Facility for their help in this work, especially J. L. Selfridge and F. H. Hollander.

University of California at Los Angeles
Los Angeles, California

1. D. H. Lehmer, "An extended theory of Lucas' functions," *Ann. of Math.* v. 31, 1930, p. 419–448.

2. H. Riesel, "Mersenne numbers," *MTAC*, v. 12, 1958, p. 207–213.

3. R. M. Robinson, "Mersenne and Fermat numbers," *Proc. Amer. Math. Soc.* v. 5, 1954, p. 842–846.

4. R. M. Robinson, "A report on primes of the form $k \cdot 2^n + 1$ and on factors of Fermat numbers," *Proc. Amer. Math. Soc.* v. 9, 1958, p. 673–681.

REVIEWS AND DESCRIPTIONS OF TABLES AND BOOKS

18 [F].—Roger Osborn, *Tables of All Primitive Roots of Odd Primes Less than 1000*, University of Texas Press, Austin, 1961, 70 p., 30 cm. Price $3.00.

This slim volume lists all 28,597 primitive roots of the 167 odd primes less than 1000. These tables were computed on an IBM 650. The program and running times are not indicated. The most extensive earlier table, as noted by the author, is due to Chebyshev and extends to $p = 353$.

There also is a small table of statistical information. Perhaps the most interesting column here lists the number of (positive) primitive roots less than $p/2$ for each prime p. Of the 87 primes $\equiv -1 \pmod 4$, eight have exactly one-half of their primitive roots less than $p/2$. The seven primes 223, 379, 463, 631, 691, 883, and 907 have more than one-half less than $p/2$. The remaining 72 primes have less than one-half there. The author associates this preponderance with the well-known fact that more than one-half of the quadratic residues of such primes lie in this interval.

For the primes $\equiv +1 \pmod 4$ this column is clearly redundant, since it is easily seen that if g is a primitive root for such a prime then so is $p - g$. For these primes the real interval of interest is $p/4 < g < 3p/4$. Since the quadratic non-residues are in excess here, one would expect the primitive roots to also be preponderantly in excess, since approximately three-fourths of all non-residues are primitive roots.

D. S.

19 [I, X].—D. S. Mitrinović & R. S. Mitrinović, *Sur les nombres de Stirling et les nombres de Bernoulli de l'ordre supérieur*, *Publ. Fac Élect. Univ. Belgrade* (Série: *Math. et Phys.*), No. 43, 1960, 64 p. (French with Serbian summary.)

The tables in this paper extend those given in previous papers, especially the three reviewed in *Mathematics of Computation*, v. 15, 1961, p. 107. The notation used is explained in that review.

Table I (p. 15–44) gives $(-)^m C_m^k$ for $k = 0(1)32$, $m = 33(1)50$, and for $k = 33(1)49$, $m = k + 1(1)50$,

Table II (p. 45–50) gives S_n^{n-m} for $m = 33(1)49$, $n = m + 1(1)50$, and also for $m = 50$, $n = 51$.

Table III (p. 51–63) gives S_n^{n-m} for $m = 1(1)3$, $n = 201(1)1000$.

The tables were computed on desk machines. Checks made by the authors were supplemented by comparison with Miksa's unpublished tables and by many-figure computations made in laboratories at Liverpool, Rome, and Munich. A bibliography of 26 items is given.

A. F.

20 [K].—B. M. Bennett & P. Hsu, *Significance Tests in a* 2×2 *Contingency Table: Further Extension of Finney-Latscha Tables*, February 1961. Deposited in UMT File.

These manuscript tables constitute an extension for $A = 21(1)30$ of tables prepared by Latscha for $A = 16(1)20$, and supersede the previous tables by the present authors for $A = 21(1)25$. (See Review 9, *Math. Comp.*, v. 15, 1961, p. 88–89.) The format and precision of those tables (four decimal places) is retained in this addendum.

J. W. W.

21 [K].—Colin R. Blyth & David W. Hutchinson, *Tables of Neyman Shortest Unbiased Confidence Intervals (a) for the Binomial Parameter (b) for the Poisson Parameter*, (reproduced from *Biometrika*, v. 47, p. 381–391, v. 48, p. 191–194, respectively) University Press, London, 1960, 16 p., 28 cm. Price 2s. 6d.

Anscombe [1] observed that exact confidence intervals for a parameter in the distribution function of a discrete random variable could be obtained by adding to the sample value, X, of the discrete variable a randomly drawn value, Y, from the rectangular distribution on $(0, 1)$. Eudey [2] has applied this idea in the case of the binomial parameter, p, to find the Neyman shortest unbiased confidence set. The present authors use Eudey's equations for a uniformly most powerful level 1-α test of $p = p^*$ vs $p \neq p^*$ based on an X in a sample of n, which give the acceptance interval $a(p^*)$ determined by a value of Y in the form $n_0 + \gamma_0 \leqq X + Y \leqq n_1 + \gamma_1$ in which n_0 and n_1 are integers and $0 \leqq \gamma_0 \leqq 1$, $0 \leqq \gamma_1 \leqq 1$. These are solved for γ_0 and γ_1 in terms of n_0 and n_1 and the given X, n, and α. Then trial values of n_0 and n_1 are used until the resulting γ_0 and γ_1 are both on $(0, 1)$. The computation was carried out on the University of Illinois Digital Computer Laboratory's ILLIAC. The program used for arbitrary n, α prints out $n_0 + \gamma_0$, $n_1 + \gamma_1$ for any equally spaced set of p^* values. From these the Neyman shortest unbiased α-confidence set for p, $X + Y \epsilon \alpha(p^*)$ can be read off to 2D. The tables give such 95% and 99% confidence intervals for p to 2D for $n = 2(1)24(2)50$ and $X + Y = 0(.1)5.5$ for $n \leqq 10$, $0(.1)1(.2)10$ for $11 \leqq n \leqq 19$, $0(.1)1(.2)6(.5)15(1)17$ for $20 \leqq n \leqq 32$, and $0(.2)2(.5)23(1)26$ for $34 \leqq n \leqq 50$. For n, $X + Y$ not tabled, one enters the table at n, $n + 1 - (X + Y)$ and takes the reflection about $p = \frac{1}{2}$ of the interval given.

Similar confidence intervals for the Poisson parameter, λ, were found by the same method. The table gives Neyman shortest unbiased 95% confidence intervals for λ to 1D for $X + Y = .01(.01).1(.02).2(.05)1(.1)10(.2)40(.5)55(1)59$ and to the nearest integer for $X + Y = 60(1)250$. For the same values of $X + Y$, 99% confidence intervals are given to 1D for $X + Y \leqq 54$ and to the nearest integer for $X + Y > 54$.

C. C. Craig

The University of Michigan
Ann Arbor, Michigan

1. F. J. Anscombe, "The validity of comparative experiments," *J. Roy Statist. Soc. Ser. A.* v. 111, 1948, p. 181–211.
2. M. W. Eudey, *On the Treatment of a Discontinuous Random Variable*, Technical Report No. 13 (1949), Statistical Laboratory, University of California, Berkeley.

22 [L].—M. I. Zhurina & L. N. Karamazina, *Tablit͡sy funkt͡siĭ Lez͡handra* $P_{-1/2+i\tau}(x)$, Tom I (Tables of the Legendre functions $P_{-1/2+i\tau}(x)$, Vol. I), Izdatel'stov Akad. Nauk SSSR, Moscow, 1960, 320 p., 27 cm., 2700 copies. Price 34.50 (now 37.95) rubles.

This important volume belongs to the well-known series of Mathematical Tables of the Academy of Sciences of the USSR, and the tables were computed on the high-speed electronic calculator STRELA at the Computational Center of the Academy.

The Russian work has been concerned with the functions $P_{-1/2+i\tau}(x)$, where τ is real and $x > -1$. The functions are real, and satisfy the differential equation

$$(1 - x^2)u'' - 2xu' - (\tfrac{1}{4} + \tau^2)u = 0.$$

The functions occur in potential problems relating, for example, to cones and hyperboloids of revolution; they also occur in the Mehler-Fock inversion formulas [1]. The tables for $-1 < x < 1$ and $x > 1$ are given in Volumes I and II, respectively. The formulas given in the Introduction to Vol. I are limited to those which have some application in the range $-1 < x < 1$. The values were computed from

$$P_{-1/2+i\tau}(x) = F(\tfrac{1}{2} - i\tau, \tfrac{1}{2} + i\tau; 1; \tfrac{1}{2} - \tfrac{1}{2}x),$$

where $F(a,b;c;z)$ denotes the hypergeometric function, and were checked by differencing. The main table (pages 13–312) gives values of $P_{-1/2+i\tau}(x)$ to 7S for $\tau = 0(0.01)50$, $x = +0.9(-0.1)-0.9$, without differences. (It is stated that Vol. II, which the reviewer has not seen, gives values for $x = 1.1(0.1)2(0.2)5(0.5)10(10)60$.) The interval in τ has been made narrow because applications in mathematical physics frequently require integration with respect to τ. It is stated that interpolation in τ may be performed by the three-point Lagrange formula with an error not exceeding 1.6 final units; it may be added that such an error can occur in only a small part of the table. Interpolation in x is naturally more troublesome, even well away from a logarithmic singularity at $x = -1$.

An auxiliary table on pages 315–318 facilitates use of an asymptotic series for large τ; arc cos x and four coefficients which are functions of x are tabulated to 7D for $x = 0.99(-0.01)-0.90$, without differences. Values of the Bessel functions I_0 and I_1 are required to be available for use with the auxiliary table.

A useful bibliography of 16 items averages about one misprint per item in the five non-Russian titles, the most entertaining being MacRobert's well-known book on "Spherical Harmonies" and a paper by Barnes on "Veneralized Legendre Functions."

The reviewer differenced about a hundred values without finding any error. Assuming its accuracy, this must be reckoned a valuable table.

A. F.

1. A. Erdélyi et al, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953, p. 175.

23 [X].—A. Charnes & W. W. Cooper, *Management Models & Industrial Applications of Linear Programming*, v. 1, John Wiley & Sons, Inc., New York, 1961, xxiii + 471 p., 26 cm. Price $11.95.

This book is addressed to persons interested in the application of linear programming techniques to various aspects of management planning. Much of the material has been published previously by the authors in scattered journals and texts; however, this volume offers the advantage of a unified mathematical treatment of sundry topics in mathematical programming and managerial economics within the framework of adjacent-extreme-point techniques.

The earlier parts of this volume do not require mathematics beyond college algebra. The rudiments of linear programming theory and techniques are illustrated by means of simple numerical examples. An elementary machine loading problem is introduced to elucidate such concepts as linear model formulation, approximation of model types by scaling, the dual linear programming problem, and data accuracy and program sensitivity. The stepping-stone method for the classical Hitchcock transportation problem and transshipment problem are described at length. The procedure for dealing with degeneracy is also discussed. To explicate the concept of input-output analysis, a three-industry input-output model as an example of a "static, open Leontief model" is given. Feasible solutions are obtained by the Gauss elimination method.

With the exception of the transportation algorithm, a rigorous mathematical treatment of the foregoing topics are presented in the succeeding parts of this volume. Background material from the fields of matrix algebra, convex sets, and linear systems are developed and interpreted to provide an essentially self-contained account of the mathematics relevant to the managerial applications covered in the rest of the volume.

Considerable attention is devoted to Dantzig's simplex method for solving the general linear programming problem. The basic simplex algorithm is carefully explained and illustrated with the aid of numerical examples and geometrical interpretations. Additional by-products and interpretations are obtained, such as the extension of the simplex calculations for analyzing the effects of altering (a) the stipulations vector, (b) the coefficients of the objective function, and (c) the structural vectors. Also, the role of the simplex procedure as a tool for securing proofs of several important duality theorems in the field of linear inequalities is deftly portrayed.

The application of delegation models to managerial economics is first examined along the lines of T. C. Koopman's "activity analysis models." A major purpose of such models is the determination of rules which might be applied to guide the activities of a decentralized management organization. Koopman's formulation is reduced to a series of special linear programming problems and their duals. "Efficient" solutions are obtained by the "spiral" method. Koopman's concept of "efficiency" is then generalized to provide under certain circumstances more suitable criteria for managerial applications.

Linear programming approaches to statistical problems involving inequality relationships are delineated and applied to a problem of determining an executive-compensation formula for an industrial concern. Moreover, the techniques employed to solve this problem provide an introduction to the use of adjacent-extreme-point methods to a variety of nonlinear problems encountered in management planning. Modifications of simplex criteria and procedures are developed for the case where a functional subject to linear constraints may be decomposed by linear transformations into a sum of functionals involving only a single variable. The basic shortcoming of this approach is that, in general, only a local optimum is guaranteed.

A dynamic model for production scheduling at minimum cost when the costs are unknown is solved by means of "surrogate" techniques and "subhorizon" methods. Optimizing rules are enumerated and expounded for solving an actual example for which these methods were first devised. This is followed by a proof of the optimizing properties of the rules. The effects of introducing costs, such as inventory charges, and additional constraints, such as storage limitations, are touched upon from the standpoint of possible variations in the length of subhorizons. A generalized approach to this class of problem is explored via the Kuhn-Tucker theorem for nonlinear programming.

The "classical" models of linear programming are presented with commendable clarity. Moreover, the adaptation of linear programming methods for solving nonlinear types of management problems is aptly demonstrated. However, this reviewer's enthusiasm was tempered by the fact that the present edition abounds with errors resulting from an apparent cursory attempt at editing and proofreading. This reviewer recommends that the publishers prepare an errata sheet; otherwise, the intolerable number of typographical errors will vitiate the intrinsic merits of this book as a textbook and reference.

Milton Siegel

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

24 [X].—Roman Jakobson, Editor, *Proceedings of Symposia in Applied Mathematics*, Vol. XII, "Structure of Language and its Mathematical Aspects," American Mathematical Society, Providence, 1961, vi + 279 p., 26 cm. Price $7.80.

Sponsored by the American Mathematical Society, the Association for Symbolic Logic, and the Linguistic Society of America, and cosponsored by the Institute for Defense Analyses under an Office of Naval Research contract, the symposium, held in April, 1960, included the following papers:

W. V. Quine	Logic as a Source of Syntactical Insights
Noam Chomsky	On the Notion "Rule of Grammar"
Hilary Putnam	Some Issues in the Theory of Grammar
Henry Hiż	Congrammaticality, Batteries of Transformations and Grammatical Categories
Nelson Goodman	Graphs for Linguistics
Haskell B. Curry	Some Logical Aspects of Grammatical Structure
Yuen Ren Chao	Graphic and Phonetic Aspects of Linguistic and Mathematical Symbols
Murray Eden	On the Formalization of Handwriting
Morris Halle	On the Role of Simplicity in Linguistic Descriptions
Robert Abernathy	The Problem of Linguistic Equivalence
Hans G. Herzberger	The Joints of English
Anthony G. Oettinger	Automatic Syntactic Analysis and the Pushdown Store
Victor H. Yngve	The Depth Hypothesis
Gordon E. Peterson and Frank Harary	Foundations in Phonemic Theory
Joachim Lambek	On the Calculus of Syntactic Types
H. A. Gleason, Jr.	Genetic Relationship Among Languages
Benoit Mandelbrot	On the Theory of Word Frequencies and on Related Markovian Models of Discourse
Charles F. Hockett	Grammar for the Hearer
Rulon Wells	A Measure of Subjective Information
Roman Jakobson	Linguistics and Communication Theory

Some of the authors are concerned with preformal questions, i.e., with a discursive characterization of the substance of language; Quine, Putnam, Chao, Herzberger, and Jakobson seem to have such interests. Others are fully engaged with the construction of formal systems: Chomsky, Hiż, Curry, Halle, Abernathy, Peterson and Harary, Lambek, Mandelbrot, and Wells. Oettinger, Yngve, and Hockett aim at description of linguistic processors—natural or artificial—rather than at characterizations of language, although all three have formalisms to display. Eden, working on handwriting, might be placed with one of the latter two groups. Goodman's contribution is the exposition of a branch of mathematics in its potential application to linguistic theory. Gleason shows the application of classification theory to a major branch of linguistics, the tracing of historical connections among languages.

A cursory inspection of this volume would suggest that the "structure of language" is just its grammatical—or, more narrowly, syntactic—structure. Mandelbrot objects to the identification of "linguistics" and "grammar" (pp. 211–214), but mathematical formalization of linguistic theory is going forward more rapidly in syntax than in any other area, and it is, as Jakobson remarks (p. vi), mathematical logic and the theory of recursive functions in particular that is being applied. Mandelbrot seems to agree with his opponents that "statistical" and "grammatical" models are "contradictory." He supposes that they must remain so; a different possibility is that grammatical models will furnish a structure on which statistical models can be developed. Grammar in any case is not the whole of linguistics, and problems like Gleason's will probably be brought to computing centers more often in the future.

Computational linguistics has been hampered by lack of sufficient and sufficiently sound publications in mathematical linguistics; this volume should be studied by any linguist or mathematician who proposes to program syntactic operations, whether for research purposes or in connection with such applications as machine translation.

David G. Hays

The RAND Corporation
Santa Monica, California

25 [Z].—Donald P. Eckman, Editor, *Systems: Research & Design*, John Wiley & Sons, Inc., New York, 1961, xiii + 310 p., 23 cm. Price $8.50.

This book is the Proceedings of the First Systems Symposium at Case Institute of Technology. It contains a Foreword, a Preface, and fourteen papers concerning systems research and systems design. The fourteen papers vary in style, most noticeably with regard to bibliographic reference. Some are simply advice from the author without reference to other work, others have extensive bibliographies. Only one pertains directly to the mathematics of computation, "A problem in the design of large-scale digital computer systems" by R. J. Nelson. This paper is devoted almost entirely to the problem of designing a machine which would be efficient in selecting the largest number of a set and (by implication) in other sorting problems. No specific design is arrived at, but a facility for scanning a region of the memory is suggested; the ideas may mislead some readers if they are unfamiliar with threshold search commands such as that of the Control Data Corporation 1604 computer and with the engineering details of comparison circuits.

Other papers have implications connected with the mathematics of computation, as would be expected in any current book on large systems. Thus in the Foreword, Simon Ramo remarks that "it could be said that systems engineering in today's sense became possible only with the introduction of the large digital computer." However, the papers in this volume contribute few direct suggestions concerning this use, and concern themselves largely with other general and specific aspects of systems engineering.

C. B. Tompkins

University of California
Los Angeles, California

26 [Z].—Daniel D. McCracken, *A Guide to FORTRAN Programming*, John Wiley & Sons, Inc., New York, 1961, viii + 88 p., 28 cm. Price $2.95.

The usefulness of Fortran as an automatic programming system available on many different computers has prompted Dr. McCracken to publish this guide. It is addressed to people who have no programming experience but have a requirement to accomplish scientific computation or wish to get some appreciation of how this can be done.

The guide is developed pedagogically, with numerous examples, and includes a set of detailed case studies which provide examples from several fields of effort. These case studies illustrate the essential features of Fortran and suggest the range of its applicability.

An appendix summarizes the characteristics of a number of Fortran systems that have been established for different computers.

Abraham Sinkov

National Security Agency
Department of Defense
Washington 25, D. C.

27 [Z].—Francis J. Murray, *Mathematical Machines*, Vol. 1 and 2, Columbia University Press, New York, 1961. V. 1, vii + 300 p., 26 cm. Price $12.50. V. 2, vii + 365 p., 26 cm. Price $17.50.

Volume one of Professor Murray's two-volume work on mathematical machines, is concerned with digital computers. There are two parts in Volume 1: part I on desk calculators and punched card machines, and part II on automatic sequence digital calculators. These digital devices are presented in the order of increasing competence and complexity.

In part I, there are eight chapters. The first four chapters describe desk calculators, from the basic idea of register and counter to the description of many commercial automatic calculators. Chapter 5 covers electrical counters and accumulators. Punched card machines are presented in Chapters 6 and 7, and sequence calculators such as calculating punch and electronic calculator in Chapter 8.

Part II consists of ten chapters. The first four chapters describe the logic aspect of the computer as well as digital arithmetic. Chapter 5 is a general discussion on the use of Boolean analysis. Chapter 6 is concerned with circuit elements. The programming aspects are covered in Chapters 7, 8, and 9. Chapter 10 is a very brief survey of digital computers.

In this volume, the author succeeded in many cases in bringing out the principles and fundamental ideas. An example is the exposition on desk calculators. Although the material is mostly descriptive, it will serve a useful purpose as a general reference.

Volume two of Professor Murray's work on mathematical machines presents the subject of analog devices. There are three parts: part III on continuous computers, part IV on true analogs, and part V on mathematical instruments.

Part III consists of fifteen chapters. After a brief introduction in Chapter 1, Professor Murray describes mechanical adders, multipliers, dividers, and other mechanical components in Chapter 2. Cams, gears, and their computing applications constitute Chapter 3. This is followed by an excellent presentation on mechanical integrators, differentiators, and amplifiers in Chapter 4. Chapter 5 is a review of circuit theory. Computation by using potentiometers and condensers are described in Chapter 6, vacuum tube amplifiers in Chapter 7, electromechanical components of D'Arsonval movement, watt-hour meters, and synchros in Chapter 8, electrical multipliers including time division multipliers, strain gauge multipliers, step multipliers, cathode ray multipliers in Chapter 9, and function generation by using mechanical, electromechanical and electronic means in Chapter 10. Chapters 11 through 13 describe equation solution: linear equations in Chapter 11, harmonic analysis and polynomial equations in Chapter 12, differential equations in Chapter 13, and error analysis in Chapter 14. Chapter 15, the last chapter of this part, discusses the use of digital check solutions obtained by using numerical methods when the analog solution has narrowed down the range of parameters.

Part IV, consisting of nine chapters, presents the idea of true analogs. True analogs are direct analogies on which measurements can be taken more conveniently or more economically than the analog devices described in part III. The author examines the theory of true analogs and includes descriptions of dimension theory, models, and principles of spatial relationships. True analogs that are described include the use of electrolytic tanks, electrically conductive sheets, stretched membranes, photoelastic models, and electromechanical analogies.

Part V consists of five chapters. It deals with mathematical instruments that operate on data in a specified form and perform a few mathematical operations. These devices include slide rule, plotting devices, planimeters, integrometers, integraphs, and other geometrical and trigonometrical devices. This part is rather unique.

This volume again emphasizes principles. A significant portion describes mechanical analog devices. The treatment of analog devices in volume two is more extensive than that of digital computers.

As mentioned in the book, this work was sponsored by the Office of Naval Research. These two volumes are a contribution to the study of mathematical machines, and Columbia University Press deserves credit for an excellent printing job.

YAOHAN CHU

Electrical Engineering Department
University of Maryland
College Park, Maryland

28 [Z].—E. L. Willey, Marion Tribe, A. D'Agapeyeff, B. J. Gibbens & Michelle Clark, *Some Commercial Autocodes*, Academic Press, Inc., New York, 1961, vii + 53 p., 23 cm. Price $2.50.

Some Commerical Autocodes is a study of nine programming languages applicable to commercial data processing problems, compiled in a tabular form by language elements. The study is based upon information available in December 1960 and does not represent the final specifications for some languages which have been, or are being, implemented for the various computers.

Frances E. Holberton

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

TABLE ERRATA

308.—A. Erdélyi, W. Magnus, F. Oberhettinger & F. Tricomi, *Higher Transcendental Functions*, McGraw-Hill Book Co., Inc., New York, 1953.

The following corrections should be made in this work:

Volume I

P. 104, eq. (43); *for* $(c - a)F(c + 1)$ *read* $(c - a)zF(c + 1)$.

P. 145, eq. (24): replace italic P and Q by their roman equivalents.

P. 150, second of eqs. (13): *for* i, *read* $-i$.

Volume II

P. 321, eq. (22): *for* k', *read* k'^2; and *for* $E(\theta, k)$, *read* $E(\theta, k')$.

J. C. Cooke

Aerodynamics Department
Royal Aircraft Establishment
Farnborough, Hampshire
England

309.—Mervin E. Muller, "An inverse method for the generation of random normal deviates on large-scale computers," *MTAC*, v. 12, 1958, p. 167–174.

The following errors have been noted in Table 5, "Inverse Values for the Normal Distribution":

j	$F(x_j)$	x_j	
		reads	should read
36	0.64062 500	0.3601*3* *003*	0.3601*2* *989*
92	0.85937 500	1.0775*0* 557	1.0775*1* 557
96	0.87500 000	1.1503*5* 938	1.1503*4* 938
100	0.89062 500	1.2298*4* 876	1.2298*5* 876
102	0.89843 750	1.2726*8* 865	1.2726*9* 865
110	0.92968 750	1.4734*5* *903*	1.4734*6* *759*
116	0.95312 500	1.6759*4* *192*	1.6759*3* *973*
119	0.96484 375	1.80989 *233*	1.80989 *224*

G. Miller Clark

Ohio State University
Columbus 12, Ohio

310.—D. J. Finney, "The Fisher-Yates test of significance in 2 × 2 contingency tables," *Biometrika*, v. 35, Parts 1 and 2, May 1948.

These tables have been checked against *Tables of the Hypergeometric Probability Distribution*, by G. J. Lieberman and D. B. Owen, Stanford University Press, 1961. All the entries were found to be correct, except for the following typographical error:

p. 149 $A = 6, B = 5, a = 6$ Probability = 0.025

for 0 .015 read 1 .015.

This error is reproduced in Table 38 on page 188 of *Biometrika Tables for Statisticians*, Volume 1, by E. S. Pearson and H. O. Hartley, University Press, Cambridge, 1954.

ANNA M. GLINSKI
JOHN VAN DYKE

National Bureau of Standards
Washington 25, D. C.

311.—R. LATSCHA, "Tests of significance in a 2 × 2 contingency table: extension of Finney's table," *Biometrika*, v. 40, Parts 1 and 2, June 1953, p. 74–86.

These tables have been checked against the Lieberman-Owen *Tables of the Hypergeometric Probability Distribution*, and the following errors noted.

A	B	*a*	prob.	for	read
16	10	14	0.05	4 .018	4 .017
16	10	14	0.025	4 .018	4 .017
16	4	15	0.005	1 .001	0 .001
17	4	16	0.05	1 .011	1 .012
17	4	16	0.025	1 .011	1 .012
19	16	13	0.025	4 .013	4 .012
19	8	15	0.05	2 .013	2 .014
19	8	15	0.025	2 .013	2 .014
19	6	19	0.05	4 .050−	4 .050
20	15	17	0.005	5 .002	5 .003
20	12	19	0.05	7 .019	7 .018
20	12	19	0.025	7 .019	7 .018

In order to be consistent with the method of construction for this table, in which the value of *b* recorded is the greatest significant value for which the corresponding probability is less than *or equal to* the probability shown at the head of the column, the following additional line should be inserted in the appropriate place in the table:

				Probability		
A	B	*a*	0.05	0.025	0.01	0.005
19	1	19	0 .050	---	----	---

ANNA M. GLINSKI
JOHN VAN DYKE

Corrigenda

ANDRES ZAVROTSKY, "Construccion de una escala continua de las operaciones aritmeticas," *Math. Comp.*, Review 63, v. 15, 1961, p. 299–300.

On page 300, line 7, *instead of* $L^n x = H(Gx - 1)$, *read* $L^n x = H(Gx - n)$.

R. T. Ostrowski & K. D. Van Duren, "On a theorem of Mann on latin squares," *Math. Comp.*, v. 15, 1961, p. 293–295.

On page 294, line 18 from the bottom, *for* $\frac{1}{4}\left(\frac{10}{5}\right)^2 = 15{,}876$, *read* $\frac{1}{4}\binom{10}{5}^2 = 15{,}876$.

Arnold N. Lowan, "On th numerical treatment of heat conduction problems with mixed boundary conditions," *Math. Comp.*, v. 14, 1960, p. 266–270.

For equations (13), (14), and (15) on page 269, read

$$T_{h,1,n+1} = \alpha T_{h-1,1,n} + (1 - 2\alpha - \beta) T_{h,1,n} + \alpha T_{h+1,1,n} + \beta T_{h,2,n} + U_{h,1,n} \qquad c_1/\Delta x \leqq h < M \tag{13}$$

$$\begin{aligned} T_{M,k,n+1} = {} & \beta T_{M,k-1,n} + \alpha T_{M-1,k,n} + (1 - \alpha - 2\beta) T_{M,k,n} + \beta T_{M,k+1,n} \\ & + U_{M,k,n} \qquad 1 < k < N \end{aligned} \tag{14}$$

$$\begin{aligned} T_{h,N,n+1} = {} & \beta T_{h,N-1,n} + \alpha T_{h-1,N,n} + (1 - 2\alpha - \beta) T_{h,N,n} \\ & + \alpha T_{h+1,N,n} + U_{h,N,n} \qquad c_2/\Delta x \leqq h < M \end{aligned} \tag{15}$$

where $U_{h,1,n}$ and $U_{M,k,n}$ and $U_{h,N,n}$ are the same as previously given. In addition, for points bounded on two sides by heat fluxes, the equations must be further modified to give

$$\begin{aligned} T_{M,1,n+1} = {} & \alpha T_{M-1,1,n} + (1 - \alpha - \beta) T_{M,1,n} + \beta T_{M,k+1,n} + U_{h,1,n} \\ & + U_{M,k,n} \quad \text{for} \quad h = M, \qquad k = 1 \end{aligned}$$

and

$$\begin{aligned} T_{M,N,n+1} = {} & \beta T_{M,N-1,n} + \alpha T_{M-1,N,n} + (1 - \alpha - \beta) T_{M,N,n} + U_{M,k,n} \\ & + U_{h,N,n} \quad \text{for} \quad h = M, \qquad k = N \end{aligned}$$

Norman J. McCormick

Ann Arbor, Michigan

A.

B.

C.

D

E.

F.

G.

H

I.

J.

K

L.

M

N.

O.

P.

Q.

R.

S.

T.

U.

V.

W

X.

Z.

CLASSIFICATION OF REVIEWS

A. Arithmetical Tables, Mathematical Constants

B. Powers

C. Logarithms

D. Circular Functions

E. Hyperbolic and Exponential Functions

F. Theory of Numbers

G. Higher Algebra

H. Numerical Solution of Equations

I. Finite Differences, Interpolation

J. Summation of Series

K. Statistics

L. Higher Mathematical Functions

M. Integrals

N. Interest and Investment

O. Actuarial Science

P. Engineering

Q. Astronomy

R. Geodesy

S. Physics, Geophysics, Crystallography

T. Chemistry

U. Navigation

V. Aerodynamics, Hydrodynamics, Ballistics

W. Economics and Social Sciences

X. Numerical Analysis and Applied Mathematics

Z. Calculating Machines and Mechanical Computation

Mathematics of Computation

TABLE OF CONTENTS

APRIL 1962

Published Quarterly for the

National Academy of Sciences—National Research Council

By the

American Mathematical Society